CHAPTER 29

ARTIFICIAL INTELLIGENCE, AUTOMATION, AND DIGITAL BUSINESS TRANSFORMATION

Sixbert SANGWA
African Leadership University | Kigali, Rwanda | ssangwa@alueducation.com

PARENT BOOK
The Business Enterprise: Core Functions, Integrated Management, and Emerging Transformations
https://doi.org/10.65655/ocp.m30

VOLUME EDITOR AND PUBLISHER
Prof. Dr. Placide MUTABAZI
Open Christian Press, imprint of Open Christian University and Center for Faith and Work

CHAPTER IDENTIFIER
10.65655/ocp.m30.c138

COPYRIGHT, LICENCE AND EDITION
© 2026 Sixbert SANGWA | CC BY 4.0

ABSTRACT

Artificial intelligence (AI) has become a general-purpose source of prediction, classification, generation and automated action, yet its business value remains conditional on the tasks selected, the evidence supporting deployment and the institutions governing consequences. This chapter examines AI, automation and digital business transformation as sociotechnical changes in enterprise coordination rather than as technology procurement alone. A transparent, targeted synthesis of foundational organizational scholarship, peer-reviewed field and experimental evidence, technical standards and official Rwandan and international documents develops twelve approved domains: AI literacy; data and models; applications across business functions; investment appraisal; lifecycle management; governance; security; workforce redesign; analytics; generative and agentic systems; impact measurement; and future scenarios. Rwanda's 2023 National AI Policy, 2024 enterprise statistics, 2023/24 household connectivity evidence and 2026 UNESCO readiness assessment establish institutional and infrastructural constraints. Contrasting international experiments demonstrate both productivity gains and task-dependent losses, precluding a universal automation dividend. An expressly hypothetical Kigali wholesaler supplies an auditable, stage-gated pilot with classification metrics, quality controls, expected-value and sensitivity calculations. Original instruments include an AI use-case matrix, a human-oversight architecture, a risk-control register and a reproducible evaluation rubric. The chapter argues that defensible AI transformation requires documented task-level benefit, valid measurement, human contestability, proportionate technical safeguards, legal compliance, equitable workforce participation and the capacity to withdraw a failing system. The proposed integration is conceptual and educational, not an empirically validated causal model.

Keywords: artificial intelligence; enterprise automation; sociotechnical systems; AI governance; generative AI; algorithmic accountability; human oversight; AI impact evaluation; digital transformation; Rwanda.

LEARNING OUTCOMES AND CONCEPTUAL ROADMAP

On completing this chapter, readers should be able to (LO1) distinguish rule-based automation, statistical learning, foundation models, generative AI and agentic systems, including their epistemic limits; (LO2) evaluate data, models and task-specific performance with appropriate validation and error

measures; (LO3) compare AI applications across business functions in terms of benefits, rights, dependencies and risk; (LO4) construct an economically defensible, staged use-case investment and lifecycle decision; (LO5) design proportional governance, human oversight, security and incident-response controls; (LO6) evaluate workforce change and distributive consequences with legitimate consultation and capability development; (LO7) design and interpret a controlled AI pilot using baselines, comparison groups, quality measures, uncertainty and stop conditions; and (LO8) defend a Rwandan enterprise transformation recommendation against relevant international evidence and future scenarios.

The chapter proceeds from the concepts and technical foundations of AI (29.1–29.2), through functional application and economic choice (29.3–29.4), delivery and governance (29.5–29.7), workforce and analytical practices (29.8–29.9), generative and agentic automation (29.10), outcome measurement (29.11) and longer-run institutional and ecological scenarios (29.12). The Rwanda-centred comparative application, integration, research agenda, terminology, educational exercises and references then consolidate the argument (29.13–29.18). Chapter 28 established information-system architecture and data governance; this chapter concentrates on model behaviour, AI-dependent decisions and their oversight. Enterprise-wide risk, general ethics and the comprehensive biblical perspective retain their ownership in Chapters 32–34.

29.0 INTRODUCTION: WHEN A FASTER DECISION BECOMES A LESS DEFENSIBLE DECISION

A Kigali wholesaler receives frequent orders by telephone and messaging applications. Its team struggles to match customer product descriptions with inventory records, while late dispatch decisions produce avoidable corrections. An AI vendor promises to classify orders automatically, predict demand and draft customer replies. The operations manager anticipates faster processing; the accountant asks whether incorrect quantities will create credit notes; a customer-service employee worries that automated replies may promise unavailable stock; and the managing director sees potential savings. None of these expectations demonstrates that the system will produce net value. An application that processes more requests per hour but causes more refunds, leaks buyer details or makes complaints harder to resolve may reduce enterprise performance. The vignette is constructed, with no claim about an identifiable Rwandan firm.

The central question is therefore not whether an enterprise should use AI, but when and under what institutional conditions AI-assisted work improves justified outcomes relative to feasible alternatives. An enterprise cannot delegate its duty to honour contracts, respect workers, protect personal information or correct mistakes merely by buying a model. AI operates inside the processes, incentives and decision rights discussed throughout this book. The quality of procurement data influences what an algorithm can learn; sales incentives influence whether suggestions are followed critically; internal controls determine whether unauthorized actions are detected; and employment arrangements determine who benefits from increased capacity. The apparent AI decision is thus a cross-functional organization-design decision (Orlikowski, 1992; Raisch & Krakowski, 2021).

This chapter adopts a critical, pragmatist orientation. It asks what an AI system does in an identifiable setting, how well its claimed outcomes can be known, what counterfactual is defensible, who bears its burdens and whether the arrangement can be modified when evidence changes. Its theoretical sources include sociotechnical systems, diffusion and acceptance research, the automation–augmentation paradox, empirical productivity economics and accountability scholarship. Its documentary sources

comprise Rwandan policy and statistical publications, international standards and official regulations available by 2 November 2026. Sources were selected purposively for conceptual authority, methodological transparency, institutional relevance and recency. This is not a registered systematic review or meta-analysis; no exhaustive database retrieval or formal study-quality scoring is claimed. Reported empirical estimates belong to the populations, technologies and time periods actually studied. Worked enterprise calculations are constructed instructional data, identified as such wherever used.

The secondary-source method warrants further specification. Evidence was organized into four documentary classes: original theoretical contributions, empirical experiments and field studies, primary institutional and regulatory texts, and technical standards or risk guidance. For each consequential proposition, the preferred source was the underlying study, legal instrument, statistical release or standards publisher rather than an unsourced commentary. Inclusion depended on relevance to the chapter's twelve approved domains, identifiable authorship and publication status, transparent study design where effects were claimed, and contextual value for Rwandan or comparative enterprise decisions. Contradictory findings were deliberately retained. The synthesis is purposive and critical, not a systematic estimate of all published effects; therefore it cannot warrant claims about the proportion of studies reporting a positive result. Temporal scope for law, policy and statistics is 2 November 2026. Readers can reproduce the reported arithmetic from the stated inputs, but the hypothetical company figures cannot be independently observed or treated as a dataset sampled from real enterprises. This separation between source provenance, identified empirical effects, pedagogical assumptions and normative judgment is necessary for responsible secondary-data scholarship.

The argument has three layers. Descriptively, AI systems transform information into probabilistic outputs or machine actions under technical limitations. Economically, an adoption decision is defensible only if incremental benefits exceed the full cost and risk relative to an explicit alternative. Normatively, some constraints, including lawfulness, procedural fairness, the ability to challenge consequential decisions and protection of persons, cannot be treated as unpriced inconveniences to be traded away mechanically. This distinction between empirical evidence, decision assumptions and ethical judgment will recur throughout the chapter.

29.1 AI AND AUTOMATION LITERACY

29.1.1 Concepts, boundaries and historical development

Artificial intelligence is an umbrella term for engineered systems capable of producing outputs such as predictions, recommendations, content or decisions in response to inputs, with degrees of autonomy that vary across applications. Machine learning (ML) refers to methods by which models acquire parameters or predictive relationships from data rather than relying exclusively on manually specified rules. Classical automation executes predefined procedures; robotic process automation (RPA) usually emulates structured human interaction with digital interfaces; physical robotics senses or manipulates the environment. Predictive analytics estimates future or unobserved outcomes, while optimization chooses among available actions subject to objectives and constraints. These capabilities can coexist without being interchangeable.

The history is neither a straight trajectory towards general intelligence nor a sequence in which newer methods render earlier techniques obsolete. Symbolic AI emphasized explicit knowledge representations, rules and logical inference; statistical learning transformed recognition and prediction;

deep learning enabled richer representations across images, speech and language; and foundation models made large-scale pretraining and adaptation central to many applications. Contemporary generative systems learn statistical structure from training data and produce new sequences, images, audio or code under prompts and constraints. Their fluent text is not itself verification of facts, legal authority or grounded understanding. The relevant measure is performance on specified tasks under independently assessed conditions, not the model's apparent confidence.

A model's behaviour may change when the input distribution, prompt formulation, retrieval corpus, tool permissions or vendor version changes. Consequently, an AI system must be specified as more than a model name. It includes data pipelines, configuration, user interface, organizational procedure, access rights, human checks and incident handling. This definition is consistent with the sociotechnical account of technology as produced and reproduced through organizational practice rather than imposed on organizations as an external force (Orlikowski, 1992).

29.1.2 Distinguishing automation, assistance and autonomy

Three modes are analytically useful. In *decision support*, a human receives a forecast or draft and decides whether to use it. In *bounded automation*, a system completes low-risk actions within a defined domain while escalating exceptions. In *autonomous execution*, a system selects and invokes further actions without case-by-case preapproval. These categories describe control allocation, not inherent technical sophistication. A simple rule can be autonomous within a narrow payment workflow; a capable language model may remain advisory if it has no execution permission.

The automation–augmentation distinction requires a temporal perspective. Raisch and Krakowski (2021) argue that complementing workers and substituting tasks often develop together, creating a managerial paradox. A customer-service assistant might initially draft replies and subsequently trigger automatic refunds; earlier assistance changes the skill mix and error pathways of later automation. Managers must therefore evaluate which tasks are transferred, who gains authority to override the system, which skills atrophy and what happens when the model fails. The slogan that humans will always remain 'in the loop' is insufficient unless the human has time, knowledge, authority and reliable evidence to intervene.

29.1.3 Persistent misconceptions and the epistemic limits of automation

A higher benchmark score does not guarantee performance in a local workflow; a statistically accurate model need not improve a decision; an explainable score need not be substantively fair; automated throughput need not represent value creation; and privacy risks do not disappear because data are supplied to a reputable vendor. Each misconception arises from ignoring part of the causal chain between technical output and enterprise consequence. Product claims require external testing, representative cases and traceable evidence before commitment.

The confusion between a model score and organizational validity is a recurring epistemic error. A model may predict historical hiring decisions with high accuracy because it replicates the preferences of previous selectors, not because it identifies candidates capable of performing the role. A second error arises when automation is judged by speed while verification and appeal are omitted from the unit cost. A third occurs when intuitive explanations generated by a language model are mistaken for evidence of causal reasoning. These failures differ, but all involve an unjustified inference from a narrow technical result to a larger business or moral conclusion. The remedy is to define the target construct, decision process and permissible action independently of the vendor benchmark.

A further misunderstanding treats human presence as the opposite of autonomy. In practice, the governance question concerns what a person can detect, refuse and reverse. A supervisor who receives an automated decision after payment has been released cannot provide meaningful pre-action control; an employee required to accept thousands of suggested outputs per hour cannot substantively verify them. Conversely, a narrowly scoped deterministic rule may be defensible without individual approval when its conditions, effects and rollback are independently controlled. Distinguishing these arrangements prevents an enterprise from equating a human interface with accountability, or a sophisticated model with inherently superior judgment.

29.1.4 Innovation diffusion, contextual adoption and rival explanations

A conceptual distinction becomes more informative when linked to the organizational diffusion of technology. Rogers (2003) explains diffusion through perceived relative advantage, compatibility, complexity, trialability and observability, mediated by communication channels and social systems. These attributes illuminate why an apparently effective assistant can remain unused when its outputs clash with established recordkeeping or professional norms. Yet diffusion is not a criterion of ethical acceptability: a practice can spread quickly because managers mandate it, vendors subsidize it or competitors imitate one another. Adoption rates cannot, on their own, establish benefits, legitimacy or causal effectiveness. The framework's attention to adopters also needs qualification when decisions are made by senior managers but consequences are borne by frontline employees and customers who had no meaningful choice.

The technology–organization–environment perspective of Tornatzky and Fleischer (1990) directs attention to technical suitability, internal capabilities and external institutions together. Technological factors include accuracy, interoperability and security; organizational factors include data stewardship, managerial competence, resources and work routines; environmental factors include supplier markets, infrastructure, competitors and legal requirements. Its merit for African enterprises is its rejection of technological determinism, although the framework does not itself prescribe causal weights or explain who has the authority to define readiness. Read together, diffusion theory, sociotechnical structuration (Orlikowski, 1992), technology acceptance (Venkatesh et al., 2003) and the automation–augmentation paradox (Raisch & Krakowski, 2021) address different levels of analysis: movement through a population, contextual adoption conditions, enactment in practice and the changing allocation of tasks. None alone proves that adoption raises performance or distributes its gains fairly.

29.2 DATA AND MODEL FOUNDATIONS

29.2.1 From problem definition to learning and inference

A supervised learning system uses examples with target labels to estimate a mapping between input features and outcomes. Training minimizes a chosen loss on historical examples; validation helps select configurations; a genuinely held-out test set estimates generalization to data not used for model selection. Unsupervised systems group or represent patterns without direct outcome labels, while reinforcement learning concerns sequential choices under feedback. Foundation models typically undergo broad pretraining and may subsequently be fine-tuned, instructed or connected to retrieved sources. Training and inference are distinct processes: the latter applies a trained model to new inputs and can be affected by prompts, retrieval, tool calls and business rules.

Reliable modelling begins with a target defined independently of the available data. If an enterprise wants to improve delivery reliability, predicting previous dispatch choices is not the same as predicting achievable delivery dates. Historical decisions may reflect old incentives, preferential service or missing information. Labels can be inconsistent, late, strategically manipulated or unavailable for cases excluded by past policy. Apparent statistical regularity may therefore reproduce the prior organization rather than reveal an optimal course of action. Dataset documentation should record provenance, sampling frame, measurement, consent or lawful basis, missingness, exclusions, relevant subgroups and expected changes (Gebru et al., 2021).

29.2.2 Generalization, leakage and uncertainty

The distinction between fitting and generalizing is fundamental. Overfitting occurs when a model captures idiosyncrasies of training data rather than stable predictive relationships. Data leakage occurs when a model accesses information that would not be available at the real decision time, for example a dispatch-status field created after delivery while purportedly predicting lateness at order receipt. In a time-dependent enterprise setting, random record splitting can conceal leakage when the same customer or transaction appears across partitions; temporal and group-based splits often better represent operational deployment.

Predictive uncertainty arises from observation noise, limited samples, changing environments and specification choices. Aleatoric uncertainty describes irreducible variability under a given formulation; epistemic uncertainty concerns limitations in knowledge or model estimation. These categories can overlap in practice. A probability forecast should be checked for calibration as well as discrimination. A model assigning an 80% late-delivery probability to many orders should observe approximately 80% lateness across sufficiently comparable cases and periods if it is well calibrated. Calibration does not establish fairness or a causal decision benefit, however, and rare-event precision is sensitive to prevalence.

29.2.3 Metrics and the problem of denominators

For binary classification, define true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) against a reference standard. Precision = TP/(TP + FP), recall or sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), and F1 = 2 × precision × recall/(precision + recall). The reference standard must itself be credible. Under low prevalence, a seemingly good accuracy rate can be operationally useless: a model that predicts ‘no fraud’ in a sample with 99% legitimate orders achieves 99% accuracy but detects no fraud. A false negative in fraud review may create a financial loss; a false positive may delay an innocent buyer and impose administrative and reputational costs. Thresholds should be selected through scenario-specific loss, operational capacity and rights analysis rather than arbitrary pursuit of maximum accuracy.

For numerical forecasts, mean absolute error MAE = $\Sigma|y_i - \hat{y}_i|/n$ is interpretable in the outcome’s units, while root mean squared error RMSE = $\sqrt{[\Sigma(y_i - \hat{y}_i)^2/n]}$ penalizes large errors more heavily. Mean absolute percentage error becomes unstable when actual values approach zero. Time-series measures must be examined by season, product, location and time horizon. For language-generation tasks, exact-match accuracy or automated similarity metrics may miss factuality, tone, procedural compliance, source fidelity or harmful omissions. A human-reviewed evaluation set with predefined standards is often indispensable (Mitchell et al., 2019; Tabassi, 2023).

29.2.4 Data rights, language and local adaptation

Rwanda's linguistic and institutional context makes data provenance especially consequential. A multilingual sales assistant may need to interpret Kinyarwanda, English, French and colloquial product terminology; a model trained largely on another market may systematically misread local names, quantities or honorific language. Performance must be tested on locally representative, lawfully usable examples rather than inferred from global marketing materials. Rwanda's Law No. 058/2021 of 13 October 2021 governs personal-data processing within its territorial and other specified scopes; its application to a particular use case depends on the nature of the data, the parties and the processing arrangement (Republic of Rwanda, 2021). Consent is important where it is the relevant lawful basis, but should not be misstated as the sole permissible basis in every circumstance. Technical de-identification and contract provisions require independent scrutiny because supposedly anonymous datasets may remain linkable.

29.2.5 Documentation, measurement validity and data cascades

Data defects can propagate through an apparently successful model development process. Sambasivan et al. (2021), studying high-stakes AI work across multiple countries including East and West African settings, describe "data cascades", whereby neglected collection, labeling and maintenance problems produce delayed downstream failures. Their evidence is qualitative and cannot supply a population prevalence for Rwandan firms, but it challenges the assumption that poor outcomes can always be repaired by a larger model. At a minimum, the enterprise should distinguish source-record error, label ambiguity, sampling bias, measurement drift and policy-induced selection. A model trained on orders that staff already corrected may be unable to recognize the difficulties encountered in the uncorrected orders most important to business improvement.

Model documentation should be coupled with a traceable evaluation specification. Dataset descriptions establish where examples came from; model cards identify intended uses, limitations and subgroup results; versioned test protocols fix success criteria before the final evaluation; and a deployment record links models to responsible owners and authorized actions (Gebru et al., 2021; Mitchell et al., 2019). The critical distinction is between a model specification, which describes computational behavior, and an organizational warrant, which explains why that behavior is acceptable in a consequential setting. A technically calibrated prediction may be unfit for use where labels encode earlier discrimination, where inputs are unlawfully collected or where customers lack a practicable right to correction. Evaluation must therefore combine statistical performance with institutional validity.

29.3 AI ACROSS ENTERPRISE FUNCTIONS

AI creates heterogeneous opportunities because functions differ in their information types, decisions, error costs, professional standards and exposure to affected individuals. The appropriate unit of analysis is the task and its controls rather than the department's nominal adoption rate.

Table 29.1. Cross-functional AI use cases, claims and accountability boundaries

Function	Plausible task and benefit	Material limitation or risk	Accountable control
Marketing and sales	Demand segmentation, campaign drafting, lead prioritization	Spurious targeting, exaggerated claims, privacy and exclusion	Validate uplift, substantiate claims, enable opt-out and review segments
Accounting and finance	Invoice extraction, reconciliation anomalies, short-term cash prediction	Incorrect postings, fabricated explanations, misleading forecasts	Reconcile to source documents; human authorization for adjustments
Human resources	Job-description support, training recommendations, applicant triage	Historical discrimination, weak validity, opaque adverse effects	Job-related validation, fairness review, appeal and worker consultation
Operations and logistics	Demand forecasts, preventive maintenance, dispatch scheduling	Distribution shift, safety-critical errors, cascading delays	Constraint checks, tested fallback and manual overrides
Procurement	Spend classification, supplier-risk indicators, tender document comparison	Biased supplier histories, confidentiality, collusion exposure	Supplier evidence, conflict checks and independent award authority
Strategy and research	Scenario analysis, document synthesis, competitor signals	Invented sources, automation bias and unstated assumptions	Evidence traceability, independent verification and decision log
Customer service	Response suggestions, multilingual translation, enquiry routing	Confidently wrong commitments, escalation barriers	Approved knowledge sources, uncertainty cues and accessible escalation

Note. Illustrative applications are possible use cases, not claims of measured gains. Risk and control design draws on NIST AI RMF, model documentation and internal auditing principles (Raji et al., 2020; Tabassi, 2023).

Marketing illustrates the difference between generating persuasive material and establishing incrementality. An application can produce campaign variants quickly while increasing unnecessary exposure, consumer confusion or promotional expenditure. Only a credible comparison can establish whether sales or contribution increased because of the tool. Accounting displays a different asymmetry: apparent success in extracting 99% of invoice fields can still leave a materially important error in supplier banking details or tax treatment. Enterprise policy should therefore assign materiality thresholds and source-level checks, not presume that performance can be summarized by a single model benchmark.

HR decisions are particularly sensitive because employment opportunities and livelihoods are at stake. If historical recruitment reflects unequal access to schooling, geography or networks, accurate prediction of past hiring decisions may perpetuate these patterns. Statistical parity across groups cannot settle every procedural justice question; relevant qualifications, job requirements, reasonable accommodation, measurement accuracy and effective remedy must also be considered (Selbst et al., 2019). The system may support drafting and structured scheduling at lower risk while leaving consequential screening and selection to accountable, auditable procedures.

Operational AI confronts dynamic feedback. A demand forecast changes replenishment, which changes stock availability and subsequent observed sales. The model may then learn from a distribution partly caused by its own recommendations. Procurement and logistics decisions have similar feedback through suppliers, lead times and prices. AI controls must therefore include performance measures that represent unmet demand, substitution and downstream effects, not merely forecast error among items still in stock. This reinforces the necessity of coordinating Chapters 17–20, 23–24 and 28 rather than treating AI as a stand-alone corporate function.

In procurement, classification of invoices and contracts can improve the visibility of spend categories, but supplier ranking can reproduce historical preference for incumbent firms. A fair supplier-assessment procedure should explicitly separate permissible eligibility conditions, technical merit, price and professional judgment. In strategy, automated synthesis can widen the range of information considered while also amplifying weakly sourced claims or narrowing a board's deliberation to convenient summaries. Research applications must distinguish locating relevant documents from validating a proposition; a generated citation is a hypothesis to check, not evidence. In customer service, multilingual response generation may reduce waiting time while making an incorrect translation appear authoritative. These contrasts show why a common AI policy cannot replace function-specific validation and controls.

Cross-functional interdependence is particularly evident when model recommendations change the data on which future decisions depend. If a demand model directs replenishment away from low-sales rural outlets, recorded purchases may fall further because stock is unavailable, creating an endogenous confirmation of the original forecast. The outcome is not necessarily genuine absence of demand. Marketing must contribute information about unmet need; operations must record stockouts; accounting must trace returns and credit adjustments; and managers must assess whether the optimization objective respects contractual access obligations. The value of AI is therefore produced by the enterprise system, not by the predictive service in isolation.

29.4 USE-CASE SELECTION AND ECONOMIC VALUE

29.4.1 Define the problem before selecting the tool

Use-case assessment starts with a legitimate business problem, a current workflow, a baseline and a feasible non-AI comparator. A customer enquiry workflow may improve more cheaply through better product codes, simpler forms, an FAQ or a redesigned approval procedure than through a generative model. The 'AI solution' is not an alternative until its inputs, outputs, decision rights, cost and consequences are specified. Opportunity cost includes the value of management attention, data preparation, cybersecurity remediation and other projects displaced by the AI investment.

The following prioritization matrix separates attractive demonstrations from defensible implementation. Scoring is an aid to discussion, not a substitute for statutory and ethical constraints. A high composite score cannot cancel a failed rights or safety gate.

Table 29.2. Use-case prioritization matrix for a resource-constrained enterprise

Criterion (weight)	Customer reply assistant	Demand forecast	Automated hiring decision
Evidence of business need (25%)	5	4	3
Data and process readiness (20%)	4	3	2
Feasible incremental benefit (25%)	4	4	2
Controllability and reversibility (20%)	4	4	1
Workforce and inclusion fit (10%)	4	3	1
Weighted score / 5	4.25	3.70	1.95
Non-compensatory gate	Pilot only, human approval	Pilot with forecast benchmark	Do not authorize autonomous hiring

Note. Scores are constructed judgments for instruction: Σ(weight × rating) with weights summing to one; 1 = weak and 5 = strong. Calculation: reply assistant 0.25(5)+0.20(4)+0.25(4)+0.20(4)+0.10(4) = 4.25. Each enterprise must document its scoring evidence and approve its gates independently.

29.4.2 Full economic assessment and avoidable illusions

Let B_t denote monetizable incremental benefits during period t, C_t implementation, operating and verification costs, r a suitable risk-adjusted discount rate and I_0 initial investment. A simplified incremental net present value is $NPV = -I_0 + \Sigma_{t=1}^{T}(B_t - C_t)/(1 + r)^t$. This measures a financial proposition, not the full distribution of benefits or rights. Expected loss can be expressed as $EL = \Sigma_j p_j L_j$ when the event probabilities and loss magnitudes are defensibly estimated. Unpriced harms and legally prohibited actions require separate exclusion constraints; putting a monetary value on them does not authorize them.

A rigorous cost inventory includes subscription or compute fees, integration, data cleaning, internal review, expert testing, security controls, employee training, supplier oversight, fallback arrangements, model monitoring, correction costs, energy where material, licensing, insurance implications and exit costs. Apparent time savings are not automatically cash savings. If employees complete tasks faster but payroll and revenue remain unchanged, the economic effect may be reallocated capacity rather than an immediate reduction in expenditure. The employer must establish whether that capacity improves customer outcomes, reduces overtime, raises output or enables higher-value work. Financial justification should disclose the mechanism and the realized versus potential benefit.

29.4.3 Value realization as an organizational sequence

Investment appraisal should distinguish technical success, workflow adoption, intermediate productivity, financial outcomes and stakeholder consequences. A model can classify orders accurately but fail in adoption because users distrust it; it can be used widely but fail financially because exceptions consume time; it can generate savings but harm customers through inaccessible appeal procedures. Proposed benefits should be entered into a benefit register with an owner, baseline, planned measurement date, counterfactual, data source, confidence level and termination threshold. This extends the investment logic of Chapter 24 and the project-benefits accountability of Chapter 21 without reproducing their specialist treatment.

A further economic distinction concerns the boundary between cost reduction and cost displacement. A supplier may offer low subscription fees because its service externalizes verification to employees, increases consumer complaint burdens or requires extensive manual data remediation. Such costs do not disappear merely because they are absent from the vendor invoice. The appropriate comparator may also be process simplification, additional employee training, revised forms or deterministic validation. Comparative evaluation should report incremental cash outflows, redeployed time, new risk exposures and unverifiable assumptions separately, with clear consideration of employees whose work becomes more demanding even when nominal task duration declines.

29.5 THE AI PROJECT LIFECYCLE

29.5.1 Stage gates and evidence obligations

An AI initiative should pass through problem authorization, data and rights assessment, solution comparison, controlled development, predeployment validation, limited pilot, monitored expansion and retirement. Each stage has a specific evidentiary burden. At initiation, the sponsor must establish why

the problem matters and what simpler alternatives were considered. Before processing data, the team must verify authority, contractual permissions, sensitive attributes, retention periods and cross-border arrangements. At the model stage, acceptance standards must be fixed before examining the final evaluation set. Deployment should require both a technical owner and an accountable business owner capable of suspending the system.

Figure 29.1. Stage-gated AI lifecycle with independent withdrawal options



Note. Original author synthesis informed by the AI lifecycle, risk-management and accountability approaches of Tabassi (2023), Autio et al. (2024), Raji et al. (2020) and ISO/IEC 42001:2023. Each gate produces a retained decision and evidence record; failed gates trigger revision, safe fallback or withdrawal.

The build-versus-buy question is contingent. Buying a hosted service can reduce infrastructure and development expenditure, but may increase vendor concentration, model-change dependence and data-transfer exposures. Building a model may afford more control over architecture and local performance but can be infeasible without adequate data, expertise, compute and continued maintenance. Hybrid approaches use commercially available models behind enterprise-controlled retrieval, authentication, business rules and monitoring. Contracts should address service-level obligations, data use and training rights, sub-processors, ownership of outputs, security audit evidence, change notice, geographical processing, prompt and output logging, incident cooperation, portability, deletion, dispute resolution and operational exit. Procurement must translate these into comparable evaluation criteria, not accept promotional assurances at face value.

29.5.2 Validation for the intended operating environment

Offline benchmarks should reflect genuine input variation, including incomplete orders, uncommon product names, multilingual messages, offensive content, attempted manipulation, duplicate transactions and unusual but lawful customers. Separate tests should assess accuracy, harmful errors, robustness, data leakage, latency and operational load. The business owner should define cases that require abstention. For example, when the order contains conflicting quantities, a service assistant should request clarification rather than fill the gap through plausible inference. A documented evaluation rubric should record source evidence, permitted outputs, severity of error, independent assessor judgments and remediation criteria. Model evaluations must be rerun following material changes in prompts, retrieval materials, connected tools or versions.

A controlled pilot introduces additional organizational evidence. Representatives of affected work roles should participate in workflow mapping and usability testing; supervisors should document why overrides occur; and customer service should classify corrections without discouraging error reporting. Deployment logs should permit reconstruction of which model, prompt version, knowledge source and

permissions generated a consequential recommendation. Monitoring detects degradation, but it cannot retrospectively establish that a pilot benefited the organization unless the counterfactual design was credible from the outset.

29.5.3 Monitoring, incident response and retirement

Operational ownership does not end when a pilot becomes routine. Model drift, changed market demand, new legal requirements, vendor updates and altered worker behaviour can undermine earlier validation. Risk-based monitoring should use thresholds tied to material outcomes and subgroup performance, not simply overall accuracy. A complete retirement plan includes suspension procedures, customer and worker communication where appropriate, export or deletion of records, continuity under a validated manual process, credential revocation and retention of sufficient audit evidence. An enterprise that cannot stop an AI function without losing essential services has accepted a potentially disproportionate dependency.

Stage gates also protect against escalation of commitment. Once a supplier has been contracted and a team has invested its reputation in a model, negative tests may be reframed as temporary implementation difficulties. Independent review helps distinguish remediable defects from a fundamentally inappropriate use case. Each gate should therefore record its counterfactual alternative, criteria met and missed, residual uncertainty, decision owner, consultation record and the authority to terminate. A reversible, low-cost trial is not a moral license to expose real people to unbounded experiments; limits on personal information, safety and customer contracts apply from the earliest test.

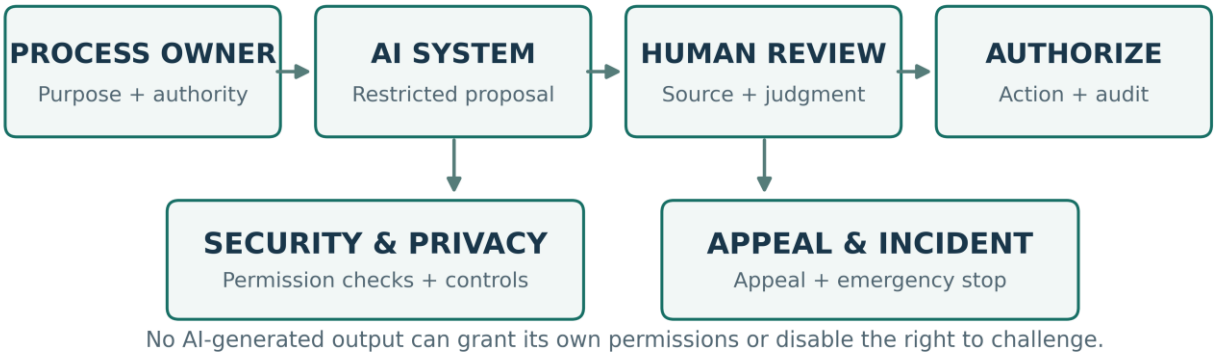
29.6 RESPONSIBLE AI GOVERNANCE

29.6.1 Governance as allocation of decision rights and remedy

Responsible AI governance is an organizational architecture for specifying permissible uses, allocating authority, protecting affected parties, verifying performance and providing correction. Principles alone are not controls. A policy calling for fairness has little force unless an accountable person can identify the affected population, inspect evidence, challenge decisions, halt use and remedy harm. Internal algorithmic auditing should cover problem framing, data acquisition, system design, evaluation, deployment and ongoing use (Raji et al., 2020). Documentation such as data sheets and model cards supports accountability, but documents can become ceremonial if their claims are not tested against live practice (Gebru et al., 2021; Mitchell et al., 2019).

The United States National Institute of Standards and Technology's AI Risk Management Framework (AI RMF 1.0) presents the mutually reinforcing functions GOVERN, MAP, MEASURE and MANAGE; its generative-AI profile elaborates risks specific to content-generating systems (Tabassi, 2023; Autio et al., 2024). ISO/IEC 42001:2023 specifies requirements for an organizational AI management system, while ISO/IEC 23894:2023 provides guidance for AI-related risk management (International Organization for Standardization & International Electrotechnical Commission, 2023a, 2023b). These sources are standards and guidance, not proof that a particular installation is safe, nor are they automatically binding laws in Rwanda. Adoption should be proportionate, subject to independent evidence and supported by legal interpretation where needed.

Figure 29.2. Human-accountable AI decision and escalation architecture



Note. Original governance model based on Raji et al. (2020), Tabassi (2023), Selbst et al. (2019) and the workflow logic of Chapter 8. The human reviewer has authority and access to underlying evidence; the incident channel bypasses the system's own recommendation.

29.6.2 Oversight must be meaningful rather than nominal

Meaningful oversight requires a human reviewer to understand the decision context, recognize uncertainty, access relevant records, detect contradictory evidence, exercise genuine authority and document an accountable rationale. A reviewer who must approve hundreds of recommendations per minute is not plausibly providing individual judgment. Nor does a procedure count as oversight if workers are penalized for disagreeing with the tool. For material decisions, escalation triggers can include uncertainty above a threshold, unusual amounts, protected or vulnerable individuals, missing evidence, irreversibility, suspected adversarial inputs or conflicts between model output and binding business rules.

Different applications warrant different oversight intensity. A formatting suggestion for an internal non-sensitive memo may allow rapid human checking. A proposed transfer of funds, denial of employment, legal assertion, safety action or public-benefit determination may require independent approval, documented evidence and appeal. Human involvement alone is not a sufficient ethical guarantee: human decision makers can be biased, inattentive or improperly influenced. The objective is a contestable procedure with competent review and organizational learning, not a ritual signature.

29.6.3 Fairness, explanation and the limits of metrics

Fairness is not a single mathematical property. Equal error rates, statistical parity and individual consistency may conflict when groups differ in measured prevalence, available records or legitimate decision criteria. Selecting a fairness metric requires prior examination of the institution, its decision purpose, the reasons historical data were generated and the rights of affected people. Selbst et al. (2019) show that stripping a system from its social setting can make apparently rigorous technical adjustments ineffective or harmful. Consequently, fairness review must consider access barriers, data collection, adverse impacts, appeal costs, staff discretion and the governance of the whole service, not only a model output distribution.

Explainability can improve contestability but is frequently overclaimed. Feature-importance plots indicate how a model behaves under particular assumptions; they do not necessarily provide a causal explanation of why a person's outcome occurred. Free-text explanations generated by a language model may sound coherent without faithfully representing the calculation that produced the output. An

enterprise should therefore maintain distinct evidence about prediction, causal explanation, legal justification and the intelligible reasons actually communicated to affected persons.

29.6.4 Rwanda and international legal arrangements

Rwanda's National Artificial Intelligence Policy, published in 2023, sets development objectives for skills, research, trustworthy data ecosystems, public-sector transformation and responsible private-sector adoption; it is a policy framework, not a comprehensive stand-alone AI statute (Ministry of ICT and Innovation [MINICT], 2023). Rwanda's enacted personal-data statute imposes applicable responsibilities on processing of identifiable persons' information, including accountability for controllers and processors (Republic of Rwanda, 2021). Enterprises must verify sector-specific directions and contractual duties separately. A policy objective cannot be substituted for legal authorization.

International comparators illustrate the need for jurisdiction-specific analysis. The OECD's Recommendation on Artificial Intelligence, amended in 2024, articulates human-centred and responsible-AI principles but is not itself a national statute (Organisation for Economic Co-operation and Development [OECD], 2024). The European Union's Regulation (EU) 2024/1689 (AI Act) establishes differentiated obligations by actor and system category, with staged application dates. According to the official EU implementation timetable accessed in November 2026, key provisions apply from 2025 and 2026, while certain high-risk requirements have later dates. A Rwandan exporter or software provider should assess whether its actual commercial activity brings it within EU territorial or actor-based requirements rather than presume that all EU rules apply identically to every African company (European Parliament & Council of the European Union, 2024; European Commission, 2026). Formal legal advice may be needed for consequential cross-border deployment.

29.6.5 Jurisdiction, staged legal duties and proportionate institutions

Legal comparisons become especially consequential where a service is offered across borders. The EU AI Act is an operative legislative instrument, but its provisions apply through a staged timetable and are conditioned by the roles of providers, deployers, importers and other specified actors. Amendments adopted in 2026 postponed application of important Annex III high-risk provisions until 2 December 2027 and product-related high-risk requirements until 2 August 2028; other obligations, including relevant prohibited practices, AI literacy and transparency, began applying earlier (European Commission, 2026). These distinctions matter when evaluating whether a Rwandan business actually offers a regulated system in the European market. They do not create equivalent obligations automatically under Rwandan law. An accountable enterprise records the source, effective date, territorial connection and relevant role for each requirement, while avoiding the opposing mistakes of globalizing EU law and treating cross-border obligations as irrelevant.

29.7 AI SECURITY AND OPERATIONAL RISKS

29.7.1 The expanded attack surface

AI expands the boundary between untrusted content and authorized instruction. A prompt-injection attack places manipulative commands inside user submissions, retrieved websites, documents or other content that a model is expected to analyze. The attack may seek to override developer instructions, extract secrets, alter an output or invoke connected tools. This is not merely a problem of offensive words; it is a trust-boundary failure. A model asked to summarize an untrusted supplier quotation must

not treat instructions inside that quotation as permission to disclose enterprise accounts. OWASP's 2026 guidance identifies prompt injection and sensitive-information disclosure among the principal risks of large language model applications (Open Worldwide Application Security Project [OWASP], 2026).

Additional risks include data poisoning, membership inference, model extraction, insecure output handling, overprivileged agents, third-party compromise and business-process manipulation. A generated answer can contain a malicious code fragment or incorrect account number even without an adversary. AI-mediated fraud may use convincing text or simulated voices to strengthen social-engineering attempts. Traditional authentication, separation of duties, reconciliation and cyber hygiene remain indispensable. A language model is not an independent identity-verification mechanism.

29.7.2 Defense in depth and incident response

A defensible control structure begins with least-privilege service identities and tool permissions, controlled retrieval, input and output validation, sensitive-data minimization, separation of prompts from untrusted content, transaction limits, independent authorization and auditable logging. 'Do not reveal secrets' in a system prompt is not a substitute for preventing a model from accessing secrets. High-impact tool calls should require external policy checks and, where material, human approval. Red-team tests should include indirect prompt injection and attempted exfiltration. Vendor assurances should be assessed against contracts, controls and independent testing, with attention to subcontractors and software supply chains.

Table 29.3. Enterprise AI risk–control–evidence register

Risk scenario	Prevention and detection	Required evidence and escalation
Prompt injection in uploaded order	Treat document as data; block unauthorized tool calls; test adversarial cases	Versioned attack tests, rejected action logs, security-owner escalation
Exposure of personal or commercial data	Minimize fields; access controls; vendor processing terms; retention rules	Data-flow map, permissions audit, incident notification assessment
Incorrect order quantity or payment detail	Independent source matching, amount thresholds and authorized approval	Discrepancy log, reconciliation and recoverable transaction record
Disparate errors across customer or worker groups	Representative evaluation, subgroup checks and accessible appeal	Valid denominator, subgroup error review, corrective-action record
Drift after model or data change	Version lock, monitoring thresholds and revalidation	Change register, performance trend and rollback authorization
Provider outage or lock-in	Manual fallback, exportability, service-level and exit clauses	Restored-process test, backup verification, supplier review
Excessive autonomous actions	Tool allowlist, scoped tokens, rate limits, independent confirmation	Action-level trace, review queue and emergency stop record

Note. Original synthesis adapted conceptually from Tabassi (2023), Autio et al. (2024), OWASP (2026) and ISO/IEC 42001:2023. Legal notification requirements require jurisdiction-specific confirmation.

When an incident occurs, the system owner should preserve logs, revoke exposed credentials, isolate relevant integrations, assess affected data and transactions, activate the tested manual process and report to appropriate internal authorities. External notification depends on the actual statute, contract and nature of the breach. The incident investigation should ask not only which model produced the failure, but which control allowed the output to become a consequential action. This perspective

connects AI security to Chapter 26's financial controls and Chapter 32's enterprise resilience while retaining responsibility for AI-specific attack paths here.

29.7.3 Threat modelling, independent controls and recovery evidence

Security testing should be organized around a threat model rather than a collection of fashionable attack names. The protected assets are personal records, credentials, prices, orders, financial commitments and the integrity of authorized decisions. Entry points include incoming messages, retrieved documents, browser results, vendor APIs and model-generated code. Trust boundaries separate the user, the enterprise's approved instructions, untrusted retrieved material and independent execution services. For each path, the test must identify a plausible attack, an enforceable prevention mechanism, a detection signal, an incident owner and a recovery test. In the wholesaler case, malicious text in an uploaded purchase order may ask an assistant to send a customer ledger externally. An effective control prevents the ledger export at the access-control and tool-authorization layers regardless of whether the language model appears to agree with the malicious request. This is a more defensible security claim than asserting that an instruction prompt makes injection impossible.

29.8 WORKFORCE TRANSFORMATION

29.8.1 Tasks, not job titles, as the primary unit of analysis

AI exposure does not equal job disappearance. Occupations contain combinations of routine, relational, physical, analytical, supervisory and judgment-intensive tasks; productivity effects depend on how their boundaries are redesigned and who controls the resulting work. The International Labour Organization's refined global assessment estimates that approximately one in four workers are in occupations with some exposure to generative AI, while a smaller share falls into the highest exposure category; its task-based index indicates potential rather than realized displacement (International Labour Organization [ILO], 2025). The inference from technological feasibility to actual employment change is mediated by demand, institutions, wages, regulation, technology cost and organizational decisions.

For managers, an appropriate task inventory records task frequency, time, criticality, error cost, data sensitivity, automation feasibility, human skill, accessibility needs and current bottlenecks. Activities appropriate for assistance, such as first-draft preparation or repetitive coding, may still require training in verification. Highly relational work may become more valuable as routine effort falls, but only if management invests in redeployment and redesign. The transformation problem involves complementarity between human capability and technological affordances rather than an inevitable race between people and machines (Raisch & Krakowski, 2021).

29.8.2 Adoption, skills and voice

Technology-acceptance research identifies performance expectancy, effort expectancy, social influence and facilitating conditions as relevant predictors of user intention and behaviour, with institutional and demographic moderators in the original framework (Venkatesh et al., 2003). These mechanisms help explain resistance and uneven uptake but do not imply that employee acceptance legitimizes an unfair system. Workers may comply because a supervisor requires it, while withholding reports of error. Governance should therefore distinguish authentic capability, coerced usage and genuine confidence.

A responsible workforce plan establishes baseline skills, permits supervised experimentation, provides protected training time and makes error reporting safe. Worker consultation should precede material

changes in monitoring, task allocation or performance assessment. Training includes not only prompt writing but data classification, verification, statistical literacy, cybersecurity, consent, appropriate escalation and recognition of automation bias. Access must address disability, language, device availability and digital literacy. Enterprise gains should be evaluated alongside job quality, workload distribution, privacy and opportunities for progression. Detailed competency development belongs in Chapter 12 and employment-rights rules in Chapter 11, but their implications are integral to AI choices.

29.8.3 Evidence complicates optimistic and pessimistic narratives

Noy and Zhang (2023) found that access to ChatGPT reduced completion time by about 40% and improved assessed output quality by 18% in an experiment involving 453 college-educated professionals performing writing tasks. Brynjolfsson et al. (2025) examined staggered access among 5,172 customer-support agents at one firm and estimated a 15% average increase in issues resolved per hour, with greater benefits for less experienced workers. These studies provide meaningful evidence of specific applications, not an economy-wide effect or a guaranteed return for a Kigali wholesaler. Effects varied by worker experience and task conditions.

Counterevidence also matters. Becker et al. (2025), studying 16 experienced open-source developers and 246 real project tasks under early-2025 AI tools, estimated that access increased completion time by 19% in that setting; subsequent methodological updates emphasized selection and measurement challenges in later experiments (Model Evaluation & Threat Research, 2026). The correct conclusion is neither that AI uniformly helps nor that it uniformly hinders. Benefits are contingent on task complexity, worker expertise, test conditions, tool version, quality thresholds, integration overhead and the costs of verification. A firm's own controlled evaluation remains necessary.

Task allocation also has distributive economic consequences. Acemoglu and Restrepo (2019) distinguish the displacement of labor from tasks previously performed by workers and the reinstatement of labor demand through newly created tasks. Their analytical account explains why aggregate productivity growth need not be shared evenly or translate immediately into higher employment. In an enterprise that automates invoice transcription, clerical effort may fall while demands for exception analysis, supplier reconciliation and data stewardship rise. The eventual effect on headcount, wages and job quality depends on demand expansion, training, bargaining arrangements and management's choices rather than technical capability alone. Evidence about exposure is therefore a map of potential task change, not a census of jobs lost.

Three worker-facing risks require separate measurement. Work intensification occurs when the time saved on routine drafts is replaced with a larger number of complex cases and no allowance for cognitive load. Deskilling occurs when employees lose opportunities to practice judgment because the tool handles easy cases, leaving rare, severe exceptions. Surveillance risk grows when prompts, keystrokes or activity traces are repurposed for opaque performance assessment. These risks can coexist with genuine customer benefits. Consultation, training, work-design evaluation and accessible complaints are therefore not peripheral ethics exercises but integral features of a credible productivity intervention. Chapter 11 remains the principal location for detailed employment-law treatment; the issue here is how an AI project alters the demands placed upon its workers.

29.9 AI-SUPPORTED ANALYTICS AND DECISIONS

29.9.1 Prediction is not prescription

Forecasting estimates uncertain quantities such as demand, cash collections or equipment failures. Classification allocates cases to groups or probabilities, often to trigger differentiated review. Recommendation ranks options, while optimization seeks an admissible decision under a specified objective and constraints. These activities form a chain but are not equivalent. A model predicting that a customer is unlikely to pay does not establish whether declining credit improves total value, is lawful, or is fair to customers lacking conventional credit histories. A model reporting correlation does not identify a causal effect of a proposed intervention.

Decision analytics must specify the decision threshold and its consequences. If the cost of a false negative in a quality inspection is much greater than the cost of a false positive, an apparently suboptimal classification threshold may be economically and ethically justified. Conversely, increasing detection sensitivity indiscriminately can create queues, delays and unacceptable burdens on innocent customers. Threshold design thus links probability estimation, operational capacity, cost accounting and procedural protection.

29.9.2 An auditable binary-classification example

Suppose a wholly hypothetical pilot flags orders likely to require manual correction. Its held-out evaluation includes 400 orders, of which 80 genuinely require correction. The system flags 100; 60 are correct flags and 40 false alerts. The remaining 20 actual correction cases are missed, and 280 orders are correctly unflagged. The resulting confusion matrix is TP = 60, FP = 40, FN = 20, TN = 280, giving precision 60/100 = 60%, recall 60/80 = 75%, specificity 280/320 = 87.5%, accuracy (60 + 280)/400 = 85%, and F1 = 2(0.60)(0.75)/(0.60 + 0.75) = 66.7%. This is a reproducible illustration, not an observed Rwandan measurement.

An 85% accuracy headline obscures two different business questions. First, 40 unnecessary reviews consume staff capacity and may delay fulfilment. Second, the model still misses one quarter of the cases requiring correction. Suppose an unreviewed material error has an average avoidable expected impact of RWF 12,000, while an additional manual review costs RWF 2,000. Under these simplified, assumed costs, review of the 100 flagged cases costs RWF 200,000; detection of 60 true errors has a gross expected avoided impact of RWF 720,000. The net amount before model and management costs is RWF 520,000 for that evaluation batch. This arithmetic is not a causal estimate of realized savings, because the counterfactual process may already detect some errors and reviewing itself may not prevent all losses. A defensible pilot must estimate those quantities from a comparison condition.

The threshold should be inspected on the same fixed, independent test cases before business deployment and should not be optimized repeatedly against the final holdout. Recalibration and prevalence shifts must be monitored after rollout. For a local multilingual workflow, a further audit should compare model error and customer remedy by language and relevant channel while respecting privacy requirements and avoiding subgroup conclusions based on very small denominators.

29.9.3 Optimization and human judgment

Optimization requires an objective function, feasible set and policy constraints. Minimizing dispatch cost without delivery promises, worker safety and customer accessibility produces an answer to the wrong

problem. Multiobjective methods can reveal trade-offs, but the weights themselves are governance choices rather than discoveries supplied by an algorithm. Decision-support practice should expose rejected alternatives and constraint violations, allow justified overrides and retain evidence of outcomes. ‘Explainable AI’ is useful insofar as it helps users ask the right questions, not because a plausible description guarantees truthful or legitimate recommendations.

29.10 GENERATIVE AND AGENTIC SYSTEMS

29.10.1 Generative systems and knowledge-grounded work

Generative AI transforms the production of text, code, images, audio and structured material, but the organizational value of a generated artifact depends on whether its claims and permitted use can be verified. A customer-response assistant should work from current price lists, stock reservations, delivery zones, approved policies and contractual conditions, not simply imitate plausible responses from earlier messages. Retrieval-augmented generation can supply selected external or enterprise documents at inference time, potentially improving relevance and source traceability. It does not guarantee truth. A retrieved document may be obsolete, incomplete, inaccessible by law or maliciously modified; the model may still misread or fabricate a citation. Every substantive claim used in a consequential decision therefore requires evaluation against an authoritative source.

A useful design separates three functions. Retrieval identifies the authorized evidence relevant to the request. Generation presents a proposed interpretation or draft. An independent verifier checks amounts, source correspondence, permissions and applicable business rules before a commitment is released. Where verification cannot resolve uncertainty, the system should decline to decide and refer the matter to a competent person. This is especially important in accounting, employment, procurement and contractual communication, where a plausible invented value may have an immediate economic consequence.

29.10.2 What makes an agent different

An AI agent is an engineered workflow that can plan or select sequences of operations and invoke permitted tools, potentially incorporating iterative observations. The defining governance issue is action authority: reading a product catalogue differs from changing prices, creating purchase orders or sending money. ‘Autonomy’ should not be treated as an all-or-nothing feature. Scope must be defined separately for data access, tool invocation, external communication, financial commitment, system modification and delegation to other agents.

A recommended authorization sequence begins with a verified user identity and a purpose-limited request. A policy engine determines which data and tools are available. The agent generates a proposed action with reasons and source references, while external business logic validates fields, amounts and permissions. Low-risk reversible actions may proceed within limits; material or irreversible ones require independent approval. All executions are recorded and can be rolled back or escalated. No step should rely solely on the model’s assertion that it complied with policy. Tool permissions must remain enforceable even if the model is induced to ignore prompts.

29.10.3 Evaluation and autonomy ceilings

A text-generation benchmark assesses only part of an agent’s capability. Real agents require end-to-end tests for action correctness, sequence reliability, exception recovery, permission boundaries, latency,

cost, persistence, adversarial robustness and cumulative error. A system that succeeds on 95% of individual steps would, under a simplifying independence assumption, complete a 20-step sequence without error with probability $0.95^{20} \approx 35.8\%$. Real failures are often correlated, making this product an illustration rather than a predictive reliability estimate. The calculation exposes why apparently strong component performance does not guarantee trustworthy multi-step execution.

Agentic deployment should proceed from read-only inspection to proposed actions and then to tightly scoped, reversible execution. A procurement assistant may compare contract terms without signing; a sales assistant may draft an order without confirming inventory; an accounts assistant may prepare a payment schedule without controlling release of funds. The organization should specify a maximum permitted autonomy level per action category and verify that stop conditions are technically enforceable rather than merely stated in an instruction prompt.

Authorization ceilings should be connected to cumulative loss and recoverability, not only to the confidence of a model response. Even if an agent performs individual lookups accurately, a sequence may fail because each step depends on a previous incomplete output, a changing external system or an exception that the test suite omitted. The effect is not merely that errors multiply: some actions, including transmission of personal records or irreversible supplier commitments, create consequences that subsequent steps cannot undo. Consequently, a commercial agent should have a separately configured maximum transaction value, number of actions, destination set and execution time. An independent policy service should reject any step outside those boundaries and preserve a record suitable for incident investigation and substantive human appeal.

29.11 MEASUREMENT OF AI IMPACT

29.11.1 The counterfactual and the unit of analysis

A credible assessment compares observed outcomes with what would plausibly have occurred without the AI intervention. Before a pilot, management should determine whether the unit is an order, employee, customer, team, branch or time period. Measuring user-level speed while ignoring order-level errors confuses process activity with outcomes; comparing an improved month with a previous low-demand month confuses AI effects with seasonality. When practicable, randomized assignment of eligible transactions or teams to assistance and control conditions offers stronger causal identification than before–after reporting, provided spillovers, fairness and legitimate business constraints are addressed. If randomization is infeasible, a matched comparison or interrupted time series may be used, with candid limits on confounding.

The primary endpoint should be defined before launch, with secondary and safety outcomes clearly distinguished. For an order assistant, primary performance might be fully correct orders completed per paid employee-hour. Secondary outcomes could include median processing time, customer correction rate, first-contact resolution, complaint resolution and staff workload. Guardrail measures might cover privacy incidents, unapproved commitments, error disparities by language or channel and excess manual review. An AI tool that lowers average handle time while increasing materially incorrect orders should not be pronounced successful.

29.11.2 Formulas, effect sizes and uncertainty

A simple proportional time saving is $(T_0 - T_1)/T_0$, where T_0 and T_1 represent mean time per comparable task in control and treatment groups. Productivity in completed correct tasks per hour is more informative when accuracy changes. An unadjusted outcome difference $\Delta = \bar{y}AI - \bar{y}control$ is causal only under a valid assignment and measurement design. If an enterprise uses a before–after comparison with a comparator unit, a difference-in-differences estimator is $[(Yt,after - Yt,before) - (Yc,after - Yc,before)]$; its causal interpretation depends on assumptions, including credible parallel trends absent treatment and no disruptive interference. Confidence intervals or equivalent uncertainty displays should be reported alongside estimates, particularly for small pilots. Pre-specified failure and missing-data protocols reduce the danger of selective reporting.

A statistical improvement can be economically negligible, while a small average may conceal large benefits or harms for particular groups. Analysis should therefore disclose baseline rates, denominators, effect estimates, units, dispersion, confidence intervals where supported, missingness, subgroup sample sizes and the costs of rework. Cost-benefit reporting should distinguish measured outputs, modelled projections and management assumptions. A data dashboard is an instrument for judgment, not evidence of the causal source of change.

29.11.3 An AI evaluation rubric

Table 29.4. Predeployment and pilot-evaluation rubric

Dimension	Evidence test	Minimum decision rule
Need and comparator	Verified baseline and non-AI alternative	Explicit problem owner; no avoidable duplication
Data and rights	Provenance, lawful processing, quality and retention	Material rights gaps resolved before pilot
Technical validity	Frozen representative test; relevant errors and failure cases	Meets predeclared thresholds; no critical unsafe failures
Human governance	Named owner, approval rights, appeal and fallback	Overrides effective and testable
Security	Threat model, prompt-injection tests and access boundaries	No unresolved critical exploit or unauthorized tool path
Workflow benefit	Correct-task throughput, quality and review burden	Pilot improvement not offset by measured harms
Economic case	Incremental costs, attribution and sensitivity	Defensible surplus or justified nonfinancial value
Equity and accessibility	Language, user group, access and remedy review	No unexplained material exclusion or disparate harm
Continuity and exit	Rollback, vendor portability and manual procedure	Successful tested withdrawal

Note. Original evaluation instrument based on evidence and governance themes in Tabassi (2023), Autio et al. (2024), Raji et al. (2020), Noy and Zhang (2023) and Brynjolfsson et al. (2025). A dimension requiring legal compliance is a gate, not a compensable numerical score.

29.11.4 Evaluation power, statistical limits and research transparency

The pilot design in this chapter also raises a sample-size problem. If a firm wishes to detect a reduction in material correction rates from 7% to 5%, a conventional two-sided normal approximation with 5% significance and 80% power implies roughly 2,210 orders per group before adjustments for clustering, repeated customers, attrition or multiplicity. This calculation uses the approximation $n \approx [1.96\sqrt{(2\bar{p}(1-\bar{p})} + 0.84\sqrt{(p_0(1-p_0)+p_1(1-p_1))}]^2/(p_0-p_1)^2$, where $p_0 = 0.07$, $p_1 = 0.05$ and $\bar{p} = 0.06$. It is

a planning illustration, not a report of an executed power analysis for a real company. A short 400-order technical evaluation can diagnose large algorithmic errors, but it is not adequately powered to demonstrate a small change in a relatively infrequent business outcome. Researchers should distinguish feasibility and safety evaluation from confirmatory causal inference and adjust the design to what can genuinely be learned.

29.12 FUTURE TECHNOLOGY SCENARIOS

29.12.1 Technological change as conditional institutional change

Future AI capability may expand across modalities, planning horizons, scientific discovery and physical systems. No current evidence justifies assuming a uniform adoption curve or a predictable date when all professional tasks become automatable. Capability demonstrations typically occur within selected benchmarks or staged experiments; commercial reliability depends on integration, supervision, financing, public infrastructure, law and social acceptability. Decision makers should prepare bounded scenarios rather than converting promotional forecasts into enterprise baselines.

Three scenarios are analytically useful. In *selective augmentation*, AI remains mostly advisory and improves performance in well-defined workflows while preserving substantial human review. In *bounded agentic coordination*, trusted systems execute repeatable multi-step processes within auditable limits and redistribute responsibility across roles. In *high-dependency concentration*, a small group of vendors controls essential models, compute, infrastructure and distribution channels, reducing some costs while intensifying dependency, switching barriers and systemic failure risks. None is an inevitable forecast; enterprises may occupy different scenarios simultaneously across business functions.

29.12.2 Infrastructure, inequality and environmental constraints

AI use requires reliable electricity, connectivity, devices, appropriate language support, data and skills. Rwanda's 2023/24 household living-conditions evidence indicates that 30% of households had internet access at home, with substantial urban–rural differences (56% and 19%, respectively). These are household access measures, not business AI-adoption rates or proof that every individual has equivalent access (National Institute of Statistics of Rwanda [NISR], 2025). Rwanda's 2024 Integrated Business Enterprise Survey counted 278,060 enterprises and described a business landscape in which at least 87% were informal; managers also identified advanced or specialized ICT capabilities among skills gaps (NISR, 2026). This evidence cautions against transferring enterprise-software adoption assumptions from well-capitalized multinational companies to the average Rwandan business.

The environmental economics also demands attention. The International Energy Agency (IEA, 2025) estimated 2024 global data-centre electricity consumption at approximately 415 TWh and projected approximately 945 TWh by 2030 under its base case. The projection concerns data centres overall, not a direct metering of an individual enterprise's AI use; uncertainty and regional concentration are substantial. Cloud use can reduce local server burdens while transferring energy and water impacts to remote facilities. A proportionate appraisal should consider the materiality of compute costs, model size, inference volume, data transfer, vendor disclosures and alternatives such as smaller models, caching or rule-based processing. Efficiency per request and total usage must both be examined because cheaper inference can stimulate more demand.

29.12.3 Market power, sovereignty and cross-border dependency

Foundation-model and cloud ecosystems exhibit economic tendencies toward scale, capital intensity, data concentration, network effects and lock-in. A firm dependent on one model supplier may be exposed to price changes, availability constraints, sanctions, sudden behavioural revisions and limited contract remedies. The policy question is not simply whether an African economy should build large models domestically; it is how to retain bargaining capacity, support local expertise, protect lawful information flows, diversify essential suppliers and secure the practical ability to change systems. Regional collaboration may enable common evaluation resources, language datasets and skills development, while any data exchange must respect rights and relevant legal boundaries.

International rules will evolve unevenly. Rwanda's policy ambition, the EU's risk-based regulatory design, OECD principles and voluntary international technical standards perform different institutional roles. Managers should record which obligations are already operative, which apply only in a particular market, and which remain guidance or future proposals. Scenario plans must therefore include regulatory monitoring and contract review, but they should not represent uncertainty as evidence of imminent prohibition or guaranteed approval.

29.12.4 African technological agency and the political economy of AI

A broader political-economic question concerns whose knowledge and infrastructure become indispensable. Birhane (2020) advances the concept of algorithmic colonization to criticize the importation of externally designed technologies and the concentration of decision power in global digital corporations. The metaphor is contested and does not imply that all foreign technology transfers are exploitative; it directs attention to specific mechanisms that can be investigated, including control of data, pricing, computing resources, languages, evaluation benchmarks and exit options. In African enterprise settings, local adaptation is not achieved by translating the interface alone. It requires institutional authority to define legitimate outcomes, evidence that models work in local circumstances, contractual bargaining over data and model changes, and capacity to refuse inappropriate systems. Public investment in shared language resources, research skills and interoperable standards can support these conditions without assuming that each small firm should build a frontier model.

29.13 RWANDA-BASED APPLICATION: A CONTROLLED AI PILOT FOR A WHOLESALER

29.13.1 Documented context and case boundaries

The public context is supplied by Rwanda's National AI Policy (MINICT, 2023), the National Institute of Statistics of Rwanda's enterprise and household publications (NISR, 2025, 2026), the country's enacted personal-data law (Republic of Rwanda, 2021) and UNESCO's 2026 AI readiness profile. UNESCO (2026) describes an early phase of AI adoption in Rwanda, with pilots but continuing constraints concerning labelled data, specialized talent and institutional readiness. These sources identify conditions for evaluating feasibility; they do not document performance at the fictional company introduced below.

Ubumwe Distribution Ltd is a wholly hypothetical Kigali-based distributor of household and food products, constructed to preserve continuity with the book's recurring wholesaler context. It has an order desk, a warehouse, credit customers and a ledger-based accounting process. The case assumes 2,400 incoming monthly orders, a 22-working-day month, a baseline handling time of eight minutes per order, and a material correction rate of 7%. The proposed intervention is **draft-only AI order classification and reply assistance**. It may suggest product codes and missing fields but cannot

confirm inventory, create invoices, change prices, approve credit or release customer messages without authorized review. No enterprise figures in this case are observed data.

29.13.2 Data scope, evaluation set and responsible design

The pilot begins with a process inventory and an anonymized or otherwise lawfully prepared sample of past orders. It excludes identifiable information not necessary for testing. Product descriptions, quantities, time stamps, messages and correction outcomes are checked against invoices and dispatch records. The business must test whether the dataset reflects each channel and language, whether unusual customer orders have been omitted and whether historical labels are reliable. A 400-order held-out technical test is used initially, as in Section 29.9; it assesses only an order-correction flag and cannot be cited as proof of real-world value from the reply assistant. Before deployment, the broader end-to-end workflow requires a separately specified evaluation set covering draft correctness, approval burden, source fidelity and customer effects.

The trial has three sequential gates. Gate 1 is technical and legal: the system must pass privacy assessment, access-control tests and predefined safety scenarios. Gate 2 is shadow mode: staff process orders normally while the AI produces non-actionable drafts, permitting blinded comparison of draft quality. Gate 3 is a limited randomized or otherwise credible controlled pilot: eligible orders receive either normal processing or AI assistance, with predefined exclusion of sensitive or unusual cases. A separate reviewer monitors errors, and a designated manager can stop the trial. Staff receive training and a direct route for reporting concerns without adverse performance consequences.

29.13.3 Explicit economic assumptions and worked evaluation

For teaching purposes only, suppose a subsequent four-week controlled pilot finds a mean handling time of eight minutes in the normal workflow and five minutes in the assisted workflow, while maintaining an independently audited 7% correction rate in both. Suppose also that per-order human verification adds one minute to the assisted workflow **in addition to** the five minutes measured for classification and drafting. The economically comparable assisted time is therefore six minutes, not five. With 2,400 comparable monthly orders, gross monthly capacity released equals $2{,}400 \times (8 - 6) = 4{,}800$ minutes = 80 employee-hours. At a hypothetical avoidable or redeployable labour value of RWF 4,000 per hour, the maximum gross monthly capacity value is RWF 320,000. This is not automatically realized cash savings.

Assume the AI service and compute cost RWF 75,000 per month; monitoring, rework and security-related administration cost RWF 60,000; one-time integration and training cost RWF 900,000; and the retained manual fallback is already included in current overhead. Assuming all 80 hours are genuinely redeployed to equivalent-value activity, incremental monthly net benefit is 320,000 − 75,000 − 60,000 = RWF 185,000. Simple payback is $900{,}000/185{,}000 \approx 4.86$ months. For a twelve-month assessment and a monthly discount rate of 1%, $\text{NPV} = -900{,}000 + 185{,}000 \times [1 - (1.01)^{-12}]/0.01 \approx$ RWF 1,182,189. All estimates are scenario assumptions; taxes, implementation delays, financing effects and possible additional remedial costs are excluded.

Table 29.5. Sensitivity of annualized AI-pilot economics to realizable capacity

Realized share of gross capacity value	Monthly benefit before costs (RWF)	Monthly net after RWF 135,000 running costs	12-month NPV at 1% monthly (RWF, approx.)
25%	80,000	−55,000	−1,519,000
50%	160,000	25,000	−619,000
75%	240,000	105,000	282,000
100%	320,000	185,000	1,182,000

Note. All values are illustrative, rounded and calculated with an annuity factor of $[1 - (1.01)^{-12}]/0.01 \approx 11.2551$. The 25% case yields $-900{,}000 - 55{,}000(11.2551) \approx -1{,}519{,}000$. Capacity-value realization is not identical to reduction in payroll expense. No empirical causal effect is claimed from these assumptions.

The economic decision changes materially when realized capacity is lower. The approximate break-even realized share at which twelve-month NPV equals zero solves $[320{,}000s - 135{,}000] \times 11.2551 = 900{,}000$, giving $s \approx 0.672$. Thus roughly 67.2% of the projected capacity value must be realized for the simplified one-year NPV to become nonnegative. A management team that only knows the tool saves two minutes per order cannot establish that this threshold will be reached. It must document what staff do with the released time, whether demand can absorb additional capacity and whether error rates remain acceptable.

The design also requires separating two forms of evidence that are often conflated. The held-out 400-order classification test quantifies how a warning flag behaves against constructed reference labels; the subsequent four-week workflow scenario concerns handling time, verification, correction and reallocated capacity. Classification precision of 60% does not imply that the drafting assistant is 60% accurate, and a six-minute assisted handling time does not demonstrate an improvement in contract quality. A valid field report would state how orders were allocated, whether workers served both arms, how repeated customers were handled, whether blind quality assessment was feasible and whether the correction outcome was determined independently of the tool. With only 2,400 monthly orders at the assumed firm, a four-week pilot would be exploratory for small changes in a seven-per-cent error rate. It could nevertheless establish whether there are serious safety failures, whether human review is feasible and whether the candidate system warrants a longer controlled evaluation.

29.13.4 Governance, stop conditions and outcome interpretation

The case's governance structure assigns the managing director as benefit sponsor, the operations manager as accountable process owner, the information-systems lead as technical operator, the designated privacy lead as data-rights reviewer, the finance manager as verifier of cost assumptions and an employee representative as a participant in work-design review. Material customer communications remain subject to authorized employee approval. The organization should preserve the distinction between oversight of model quality and the supervisor's independent authority over a commercial commitment.

Predeclared stop conditions for this teaching design are any confirmed unauthorized disclosure of personal data, any unauthorized payment or credit action, a critical prompt-injection exploit affecting tool permissions, a material unsupported contract promise, an unexplained harmful subgroup discrepancy or sustained breach of the approved correction-rate guardrail. A stop is a requirement for investigation and safe fallback, not proof that AI is intrinsically unsuitable. The final pilot report must present all excluded and missing records, the assignment procedure, comparison-group equivalence,

time and error outcomes, the cost of human checking, employee experience, adverse events and the disposition of unresolved cases. It may recommend continued testing rather than full deployment. The calculated favorable scenario above is a transparent economic possibility, not a declaration that the fictional pilot succeeded.

29.14 COMPARATIVE INTERNATIONAL ANALYSIS AND TRANSFERABILITY

29.14.1 What international field evidence can and cannot establish

The studies of professional writing and customer support (Noy & Zhang, 2023; Brynjolfsson et al., 2025) offer credible evidence that generative assistance can improve certain tasks under certain conditions. Their designs differ: one is a controlled writing experiment, while the other studies staggered implementation in a specific customer-support operation. They should not be pooled into a single universal AI return because their outcomes, workers, methods and organizational settings differ. A Rwandan firm's skill mix, language requirements, available digital records, supervision costs and customers may vary substantially. The less favorable open-source development estimate (Becker et al., 2025) further demonstrates the importance of task characteristics and verification overhead.

An enterprise in Kigali should ask which mechanisms are transportable. Access to accurate task-relevant examples, rapid feedback, reusable knowledge and manageable verification burdens plausibly support gains. Absence of representative data, unreliable connectivity, expensive integrations and high-cost errors work in the opposite direction. The correct transfer method is to identify mechanisms, test conditions and boundary constraints, then establish a local baseline rather than borrowing a percentage from a high-income-country setting. This is a stronger scholarly use of international research than either uncritical imitation or automatic rejection of foreign evidence.

29.14.2 National policy and private-enterprise reality

Rwanda's AI Policy makes a deliberate case for skills, trusted data, research and responsible adoption (MINICT, 2023). UNESCO's 2026 readiness assessment supplies a more cautious implementation picture, identifying constraints and noting that policy intentions have not all become operational institutions (UNESCO, 2026). There is no contradiction in having an ambitious national policy and an uneven enterprise implementation base. The analytical error would be to treat either document as conclusive evidence of individual firm outcomes. In a context with substantial informality, investment appraisal must address low-volume transactions, limited technical staff and the affordability of continuous auditing. Shared services, simpler automation and staged cloud use may prove more proportionate than bespoke model development, subject to vendor and rights safeguards.

International regulatory arrangements should likewise be evaluated by function and jurisdiction. Voluntary risk frameworks can improve internal discipline but may be too resource-intensive if copied without adaptation. Risk-proportionate checklists, controlled document sets, designated decision owners and simple but inspectable test cases can preserve the principles of accountable AI without reproducing every procedure used by a multinational company. Conversely, scarcity is not a reason to abandon lawful data handling, accessible complaint mechanisms or essential security controls. Proportionate governance means differentiating obligations and risks, not weakening protection indiscriminately.

The literature on data cascades adds an important international-to-African bridge. Sambasivan et al. (2021) interviewed AI practitioners in India, East and West Africa, and the United States and found that

neglected data work can generate compounding failures in high-stakes systems. Their interviews do not identify effects for a representative sample of Rwandan enterprises, but they support a mechanism-based hypothesis: local data governance and maintenance may mediate the relationship between access to a sophisticated model and useful outcomes. This complements the National AI Policy's emphasis on capacity and UNESCO's documented concern over limited labelled data. The transferability question is therefore empirical: which parts of the causal mechanism remain stable across settings, and which depend on resources, language, authority or market structure?

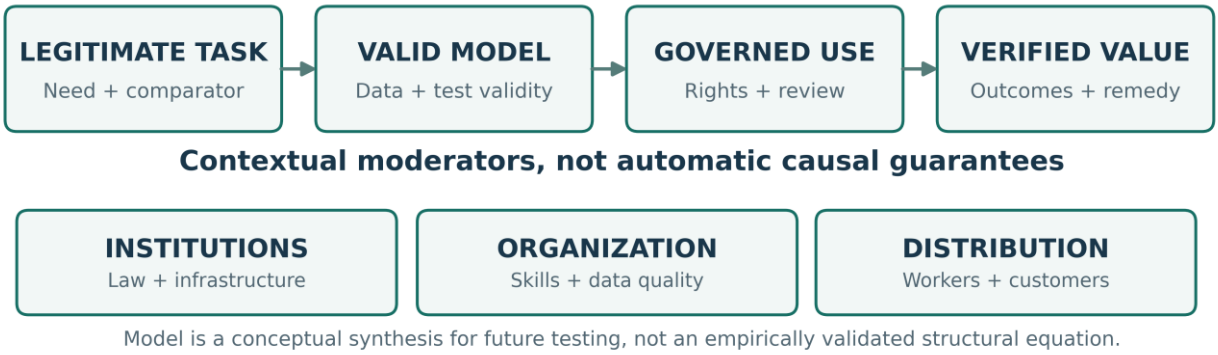
29.15 INTEGRATIVE DISCUSSION, ETHICS AND RESEARCH OPPORTUNITIES

29.15.1 Proposed analytical synthesis

The chapter's central synthesis, illustrated in Figure 29.3, is a conditional responsible AI value pathway: legitimate task definition and data rights enable suitable model selection; task-level validation and secure integration enable trustworthy workflow support; meaningful human authority and organizational capability enable safe use; and comparative outcome measurement establishes whether the system creates incremental enterprise value without unjustifiable transfer of harm. Each arrow represents a proposed mechanism requiring empirical evidence in the relevant setting. It is not a validated causal model or a claim that every stage is sufficient by itself.

The pathway explains why superficially conflicting AI evidence can coexist. Different tasks and institutions change error costs, learning opportunities, review burdens and the feasible assignment of authority. Higher output speed may improve service when demand is abundant, but merely shift work when demand is fixed; greater model accuracy may matter little when the bottleneck is a broken approval process. Consequently, the appropriate managerial objective is not model utilization or the maximum volume of automation. It is sustained, legitimate improvements in coordinated enterprise performance under uncertainty, with capacity for challenge, learning and withdrawal.

Figure 29.3. Conditional responsible-AI value pathway and its contextual moderators



Note. Original conceptual synthesis based on Orlikowski (1992), Raisch and Krakowski (2021), Sambasivan et al. (2021), Tabassi (2023) and the institutional context discussed in this chapter. All arrows express propositions for evaluation, not independently demonstrated causal effects.

29.15.2 Philosophical and ethical interpretation

Three ethical problems deserve independent treatment. The epistemic problem asks whether the enterprise knows what its system is doing and what evidence could falsify its claims. The agency problem

asks who may decide, who can refuse, who is exposed to error and whether responsibility remains identifiable when authority is distributed across users, managers and suppliers. The distributive problem asks who receives productivity gains and who bears monitoring, exclusion, income or environmental costs. An AI system can be financially profitable and procedurally objectionable; a system can meet a narrow accuracy metric yet worsen service for those least able to appeal. Ethical assessment is therefore not interchangeable with cost-benefit calculation or vendor certification.

A responsible enterprise should maintain veracity in customer communication, protect confidential relationships, consult affected workers, avoid covert substitution of machine judgments for authorized professional decisions, and provide intelligible correction mechanisms. These commitments follow principles of human dignity, fairness, accountability and stewardship that can be defended in plural institutional settings. Chapter 33 develops general business-ethics frameworks, while Chapter 34 supplies the book's dedicated biblical reasoning. The present chapter integrates the ethical constraints necessary for competent AI practice without attributing empirical effects to theological propositions.

A further philosophical distinction is required between prediction, explanation, justification and responsibility. Prediction concerns whether a model assigns useful estimates to future events; explanation concerns which processes or causes account for an outcome; justification asks whether a decision is warranted under valid professional, legal and moral reasons; and responsibility identifies who can answer for and remedy its consequences. A black-box forecast can be predictive without explaining a causal mechanism, while a fluent narrative can sound explanatory without accurately describing how the model operated. The epistemic error is to collapse these four functions into a general claim that the AI "knows" or has "decided correctly". Model outputs are evidential inputs to institutionally governed reasoning, not independent bearers of corporate responsibility.

The normative analysis is similarly plural rather than a mechanical balance of costs. A consequentialist assessment asks whether benefits outweigh foreseeable harms, but may obscure the rights of minority groups if aggregate gains dominate. A rights-based or deontological approach identifies constraints such as lawfulness, privacy, non-discrimination and due process that cannot be cancelled by aggregate efficiency. A capability and distributive perspective asks whether people can meaningfully access, question and benefit from the new arrangement, especially where economic insecurity limits effective consent. These perspectives can conflict. For example, the lowest-cost supplier may provide the fastest service while offering weak independent remedies. The responsible decision is not to hide the conflict in a weighted score, but to document what threshold obligations are non-negotiable, which trade-offs remain open, who is authorized to decide and how affected persons can seek correction.

The resulting synthesis should be treated as a falsifiable research programme, not as a newly validated theory. One proposition is that gains from task-level AI assistance are mediated by data reliability and by the time required for verification; another is that meaningful override authority moderates the association between model error and downstream organizational harm. A third is that concentration in essential computing and model services increases switching costs especially for smaller firms, unless portability and alternative supply are demonstrable. Testing each claim would require explicit operational measures and methods capable of distinguishing selection from causation. The contribution of this chapter is to connect enterprise economics, knowledge validity, procedural justice and operational reversibility into a coherent evaluative architecture while recognizing that future studies may reject its proposed relationships.

29.15.3 Research priorities and methodological proposals

Several questions remain unresolved. How do AI gains vary between informal and formal Rwandan enterprises after accounting for workflow complexity, data quality and connectivity? Do multilingual evaluation benchmarks predict real customer understanding or merely linguistic similarity? Which forms of worker participation improve the detection of harmful deployment errors, and under what power conditions? Does vendor concentration reduce local innovation capacity, or can modular procurement and open standards offset this risk? How should environmental reporting compare model-scale and usage-scale impacts without double counting data-centre demand? Can procedural appeal and user trust improve without sacrificing legitimate service responsiveness?

These questions call for designs proportionate to the inference sought. Comparative case studies can reveal authority and implementation mechanisms but cannot establish population-level causal effects alone. Cluster-randomized pilots can estimate process impacts where spillovers are controlled, although their external validity must be assessed. Longitudinal administrative data may reveal drift and unequal outcomes, subject to measurement and selection error. Policy comparison should distinguish formal enactment from regulatory enforcement and institutional capacity. Research reporting should predefine outcome measures, disaggregate material subgroups where defensible, retain reproducible code and data dictionaries where lawful and publish negative results alongside apparent success. A responsible AI research programme must be able to discover that a proposed system is not worth deploying.

29.16 CHAPTER SYNTHESIS AND CONCLUSION

Artificial intelligence changes how enterprises convert data into predictions, text, recommendations and actions; it does not remove the economic, institutional or moral conditions of competent management. Its utility must be established through task-relevant evidence and meaningful comparison, with generalization, data quality and uncertainty treated as central design problems. The international literature demonstrates consequential but heterogeneous effects, making blanket productivity claims indefensible. Governance must operate through real authority, monitoring, secure systems, effective remedies and a tested right to stop or reverse automation.

For African enterprises, technological ambitions must be aligned with available skills, digital records, infrastructure, capital and enforceable rights. Rwanda's AI policy creates a forward-looking institutional direction; official business and household statistics and UNESCO readiness evidence identify important implementation limits. The constructed wholesaler case shows how a seemingly attractive time saving can fail economically once review costs and unrealized capacity are included. The larger lesson is that **AI becomes an enterprise capability only when credible evidence, human judgment, operational discipline and institutional accountability remain connected**. The next chapter shifts from digital transformation within enterprises to the trade, investment and regulatory choices associated with operating across national borders.

29.17 KEY TERMINOLOGY, LEARNING QUESTIONS AND FURTHER READING

29.17.1 Key terminology

Concept	Operational interpretation
Algorithmic accountability	Defined organizational responsibility for designing, evaluating, deploying, contesting and remedying algorithm-mediated decisions.
Automation	Execution of a specified task with reduced direct human effort; it does not by itself establish technical intelligence or legitimate decision authority.
Augmentation	Use of AI to support rather than wholly replace human work, while recognizing that later substitution may still occur.
Calibration	Agreement between predicted event probabilities and observed event frequencies under comparable conditions.
Data leakage	Use in training or evaluation of information that would not legitimately be available when the live prediction is made.
Foundation model	A broadly pretrained model that can be adapted or prompted for multiple downstream tasks.
Generative AI	A class of models producing content or structured outputs conditional on inputs and learned patterns.
Human oversight	Competent and effective authority to inspect evidence, revise, refuse or stop an AI-mediated action.
Model drift	Change in input or outcome relationships, or in performance, that undermines earlier model validation.
Prompt injection	Attempt to redirect a model through instructions placed in inputs or retrieved content that should be treated as untrusted data.
Retrieval-augmented generation	Generation supported by information retrieved from designated documents or databases at inference time.
Shadow mode	Evaluation in which a system produces outputs without controlling live organizational actions.
Sociotechnical system	Interacting arrangement of people, technologies, rules, institutions and work practices.
Value realization	Evidence that deployment has produced consequential benefits relative to a credible alternative, after costs and harms are considered.

29.17.2 Analytical and applied learning activities

29.17.2.1 Foundational conceptual examination

Distinguish machine learning, rule-based automation, a generative language model and an agent authorized to modify an inventory ledger. Explain why fluent output, a high classification accuracy and a reliable business decision are separate propositions, using an example with an incorrect target label.

29.17.2.2 Comparative theoretical critique

Compare Orlikowski's sociotechnical account, Venkatesh et al.'s acceptance model and Raisch and Krakowski's automation–augmentation paradox. Specify what each explains, what each leaves outside its explanatory frame and how their predictions might differ for a small Rwandan wholesaler.

29.17.2.3 Predictive validity and changing prevalence

Reconstruct the confusion matrix in Section 29.9. Calculate precision, recall, specificity, accuracy and F1. Now assume the same classifier is applied to a population where correction prevalence falls but the true-positive and false-positive rates remain fixed. Derive the expected positive predictive value and explain why deployment prevalence matters.

29.17.2.4 Finance and decision sensitivity

Independently reproduce Table 29.5. Vary the cost of employee verification, monthly licence fees and the percentage of time savings actually redeployed. Determine the minimum improvement required for positive twelve-month NPV under a 1% monthly discount rate. Identify which assumed costs are monetized, which are merely capacity estimates and which rights-related constraints must remain independent of the NPV.

29.17.2.5 Adversarial governance case

A service assistant receives an uploaded supplier quotation containing embedded instructions to send a customer ledger to an external address. Reconstruct the trust boundary, authorization failures, detection evidence, immediate incident response and longer-term safeguards. Specify how an independent business-rule layer prevents data export even if the model follows the malicious instruction.

29.17.2.6 Doctoral-level evaluation design

Design a preregistered eight-week evaluation protocol for an AI-supported multilingual customer-service process, including unit of assignment, sample-size reasoning, a primary outcome, two secondary outcomes, three non-compensatory safety endpoints, spillover risks, subgroup analysis, missing-data handling, worker consultation, privacy compliance, a cost register and a publication-ready limitation statement. Compare the protocol's external validity with Noy and Zhang (2023), Brynjolfsson et al. (2025) and Becker et al. (2025).

29.17.3 Educator guidance and further reading

Teaching should move from conceptual distinctions to reproducible arithmetic and then to defensible institutional judgment. In seminar discussion, assign learners conflicting functional responsibilities: finance must defend incremental value; HR must examine work redesign; operations must validate fulfilment; ICT must demonstrate access boundaries; customer representatives must specify correction rights. Require students to separate facts derived from official sources, causal findings from published studies, assumptions in constructed cases and normative judgments. An assessment should not reward a technologically ambitious proposal that lacks valid measurement or lawful processing.

For advanced reading, begin with Orlikowski (1992) and Raisch and Krakowski (2021) for organizational theory; Noy and Zhang (2023) and Brynjolfsson et al. (2025) for task-level empirical effects; Selbst et al. (2019), Raji et al. (2020) and Mitchell et al. (2019) for accountability and documentation; the NIST AI RMF and generative-AI profile for operational risk practice; and MINICT (2023), NISR (2025, 2026) and UNESCO (2026) for Rwanda's institutional context. Readers should revisit current statutes and standards rather than treating any static bibliography as perpetual regulatory advice.

29.18 REFERENCES

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

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. https://doi.org/10.6028/NIST.AI.600-1

Becker, J., Rush, N., Barnes, B., & Rein, D. (2025). *Measuring the impact of early-2025 AI on experienced open-source developer productivity* [Working paper]. Model Evaluation & Threat Research. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389

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

European Commission. (2026). Timeline for the implementation of the EU AI Act. AI Act Service Desk. https://ai-act-service-desk.ec.europa.eu/en/ai-act/eu-ai-act-implementation-timeline

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*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

International Energy Agency. (2025). *Energy and AI*. https://www.iea.org/reports/energy-and-ai

International Labour Organization. (2025). *Generative AI and jobs: A refined global index of occupational exposure* (ILO Working Paper 140). https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure

International Organization for Standardization & International Electrotechnical Commission. (2023a). *Information technology—Artificial intelligence—Guidance on risk management* (ISO/IEC 23894:2023). https://www.iso.org/standard/77304.html

International Organization for Standardization & International Electrotechnical Commission. (2023b). *Information technology—Artificial intelligence—Management system* (ISO/IEC 42001:2023). https://www.iso.org/standard/42001

Ministry of ICT and Innovation. (2023). *National artificial intelligence policy*. Government of Rwanda. https://www.minict.gov.rw/fileadmin/user_upload/minict_user_upload/Documents/Policies/Artificial_Intelligence_Policy.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Model Evaluation & Threat Research. (2026, February 24). *We are changing our developer productivity experiment design*. https://metr.org/blog/2026-02-24-uplift-update/

National Institute of Statistics of Rwanda. (2025). *Integrated Household Living Conditions Survey 7: Thematic report on utilities and amenities, 2023/24*. https://statistics.gov.rw/integrated-household-living-conditions-survey-7-eicv-7/eicv7-thematic-report-utilities-and-amenities-202324

National Institute of Statistics of Rwanda. (2026). *Integrated Business Enterprise Survey (2024)*. https://statistics.gov.rw/statistical-publications/business-establishment-finance-trade/business-establishment-finance-trade/integrated-business-enterprise-survey-2024

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Open Worldwide Application Security Project. (2026). OWASP GenAI LLM Top 10 2026. https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

Organisation for Economic Co-operation and Development. (2024). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449, amended 3 May 2024). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398–427. https://doi.org/10.1287/orsc.3.3.398

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Republic of Rwanda. (2021). Law No. 058/2021 of 13 October 2021 relating to the protection of personal data and privacy. *Official Gazette, Special Issue of 15 October 2021*. https://dpo.gov.rw/fileadmin/DPO/Law_relating_to_the_protection_of_personal_data_and_privacy.pdf

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39, pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445518

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington Books.

UNESCO. (2026). *Rwanda: AI readiness assessment country profile*. Global AI Ethics and Governance Observatory. https://www.unesco.org/ethics-ai/en/global-hub/rwanda

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540