ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Agentic Artificial Intelligence and Machine Learning for Enterprise-Wide Intelligent Automation and Digital Transformation in Saudi Arabia under Vision 2030

Niranjan Tare

[Affiliation not supplied]

Published November 2026

Abstract

Enterprise automation is moving on from simple software rules to systems that can take in a request, work out what tools to use, co-ordinate a range of specialist agents and then make decisions based on the feedback they receive. That said, greater technical autonomy doesn't in itself lead to organisational change. This integrative review looks at how agentic AI, machine learning and the existing forms of process automation can be brought together to enable enterprise-wide change in Saudi Arabia as part of Vision 2030. It draws on 30 DOI identifiable publications from 2020–2025, covering the technological, organisational and Saudi specific evidence. A concept-based analytical approach is used to separate out what is actually demonstrated from what is merely proposed in terms of architecture and to assess the degree to which the findings can be applied to regulated enterprise workflows. Six interrelated dimensions are identified: process suitability, intelligent orchestration, data and infrastructure, human capability, risk governance and value measurement. The proposed architecture for accountable automation is one in which decision rights, evidence provenance and exception handling are all situated around a bounded agent execution loop. The Saudi case studies do suggest that management buy-in and digital maturity are important, but provide very little in the way of direct evidence on the scale of agentic deployment. The review therefore does not accept that agent autonomy, economic benefit and public legitimacy are equivalent. It does provide a context sensitive design framework and a testable research agenda for enterprises that are looking for real transformation that is reliable, interoperable and inclusive rather than just a series of isolated proofs of concept. Longitudinal, sector specific evaluation is required before any claim can be made about causal national outcomes.

1 Introduction

Saudi Arabia's Vision 2030 provides the strategic context in which productivity, industrial diversification and better public services are being pursued in parallel with digitally enabled institutional reform. In that context enterprise transformation cannot be about simply installing software. It's about the redesign of information flows, of authority, of accountability and of inter-organisational coordination. Verhoef et al make a distinction between digitisation and transformation, Hanelt et al make the point that technology and continuous organisational change are inextricably linked [1,2]. That distinction is important as workflow automation can speed up a broken process without in any way improving its legitimacy or its economic contribution. Transformation is therefore judged here in terms of capabilities and outcomes rather than in terms of the number of algorithms that have been deployed.

Classical RPA executes a set of pre-defined actions across applications. ML extends the range of things that can be done via pattern recognition, classification and forecasting. Agentic systems go further still with the ability to plan, invoke tools, delegate tasks and iteratively respond to environmental feedback. They could potentially link together procurement, customer service, maintenance and finance reporting etc. The problem with all of this is that the point at which a model makes a recommendation and that becomes an action which has consequences is inherently uncertain. The literature on RPA is replete with examples of the difficulties of implementation and governance [5–14], whilst the agent literature is focused on planning, interaction and collaboration [15–20]. There's still no strong empirical basis to equate advances in agent benchmarks with robust enterprise operations.

The Saudi context really brings this mismatch into focus. We're talking about transformation programmes in large organisations, public sector institutions and small businesses with very disparate levels of data maturity, technical capability and appetite for operational risk. There is a good body of work on Saudi enterprises looking at the conditions for adoption (management buy-in, technological compatibility, infrastructure, staff attitude etc) [23-27]. That said, that work is on AI adoption or digital transformation more generally and doesn't in any way establish that a multi-agent architecture will produce the same outcomes in Saudi companies. This review is very clear in the distinction it makes between the organisational evidence that already exists and the more speculative agent design recommendations that follow.

Three gaps underlie the study. The first is that technical agent surveys very rarely look at how ownership and exception handling map into enterprise governance. The second is that the existing literature on automation is very strong on process but predates the greater degree of discretion that is now afforded to language model agents. The third is that the existing Saudi adoption literature tends to look at intention or readiness rather than at the actual monitored, longitudinal outcomes of operation. A joined analytical model is required to explain when probabilistic decision support can safely become authorised process execution and how enterprise performance should be assessed in that context.

2 Aim, objectives and research questions

The aim is to critically synthesise the evidence from 2020–2025 and develop a governance-centric conceptual framework to explain how agentic AI and machine learning may be used to support intelligent automation across the enterprise in Saudi Arabia. The first objective is to clearly differentiate between agentic, predictive and rule-based automation in terms of decision authority and operational dependencies. The second is to identify the technical and organisational factors that will affect the extent to which automation can be integrated across enterprise processes. The third is to consider the governance, workforce and public interest implications in the context of Saudi institutions. The fourth is to propose a set of evaluation criteria and a research agenda for responsible adoption.

The synthesis is guided by four questions. RQ1 is concerned with the capabilities and boundaries that differentiate agentic orchestration from conventional automation. RQ2 is about the data, the processes and the human capabilities that enable coordinated enterprise-wide deployment. RQ3 is about the nature of the governance structures that will constrain and legitimise the action of the agents. RQ4 is about how Saudi organisations can tell the difference between real transformation and

just localised efficiency. These are questions about the mechanisms and the boundary conditions rather than any kind of promise of a universal maturity score.

3 Conceptual foundations

Digital transformation literature frames the uptake of technology as being embedded in strategy, in the way the organisation is arranged and in the evolving ecosystem [1-3]. A process perspective on this adds visibility to the pathway of execution: activities have owners, inputs, decision rules, dependencies and outcomes. In this view machine learning is about prediction and classification; automation software is about the execution of a pre-defined sequence; an agent is the one that selects an action from the set of permitted actions in response to an objective and observable state. Each of these components provides a different capability. If you conflate them you lose both the source of value and the locus of responsibility.

The automation-augmentation paradox provides a second theoretical foundation. Raisch and Krakowski argue that replacement and augmentation are not mutually exclusive but are in fact intertwined [4]. Take automated invoice matching as an example: it may well eliminate the need for the routine comparison of invoices, but will in turn create specialist roles in exception review, model evaluation and dispute resolution. The enterprise therefore needs to be actively managing the allocation of human judgement rather than simply going down the route of maximum automation. This view stops productivity metrics from masking the impact of the change on job quality, the nature of the skills required and the level of control.

Agentic systems are different from conventional scripts in that they have the potential to dynamically select their intermediate steps. Current surveys classify capabilities such as planning, memory, tool use and multi-agent coordination, but also report on unreliable reasoning, increasing interaction costs and difficulty in evaluation [15–20]. The use of external tools creates a very real interface between the probabilistic nature of language processing and the deterministic nature of business records. That interface is the central conceptual point of the present framework: autonomy must be bounded by permission, provenance, validation and reversible execution to the extent that there are consequences.

4 Review methodology

An integrative, conceptually based review was selected as the research question cuts across information systems, business process management, human-AI interaction and Saudi digital development. A quantitative effect size synthesis would be inappropriate given the very wide variation in interventions, units of analysis, outcome measures and study designs. An exhaustive database based systematic review is not claimed. What is claimed is a synthesis of a purposefully selected, DOI verifiable set of 30 papers published between 2020 and 2025, with a focus on the causal mechanisms and the context in which they operate.

Retrieval was driven by a combination of terms covering 'agentic AI', 'autonomous agents', 'multi-agent systems', 'machine learning', 'intelligent process automation', 'robotic process automation', 'digital transformation', 'governance', 'Saudi Arabia' and 'Vision 2030'. Publisher records, the bibliographic details of the papers and the reference networks of the papers identified were all used. Eligibility criteria were: publication date within the period in question (2020–2025), a DOI, a journal article or a peer-reviewed conference paper and relevance to at least one of the review questions. The

two papers uploaded did inform the structure and the comparative framing, but as they were published in 2026 they were not included in the 2020–2025 references.

Evidence was structured in a concept matrix covering publication type, technology mechanism, organisational context, performance construct, governance issue and relevance to Saudi enterprises. The synthesis looked at the findings both within and across the following categories: digital transformation theory; traditional process automation; agent architectures; productivity evidence; Saudi readiness and adoption; and language model risks. Empirical studies were separated out from surveys, policy discussions and conceptual work. The illustrative use cases and proposed architecture are purely analytical in nature and do not represent the outcome of any implementation by the author.

The assessment focused on research setting clarity, methodological transparency, the extent to which the evidence supported the claims, potential for confounding and transferability. A review finding was considered stronger if it was underpinned by empirical organisational data and where the outcome measures went beyond just perception. No invented screening totals, reviewer agreement scores, primary datasets or effect estimates are introduced. The limitations are that the search was targeted rather than exhaustive, there is a clear English language bias and there simply wasn't enough data available on longitudinal Saudi agentic deployments. These all serve to underpin a cautious interpretation rather than any claim of completeness.

5 Thematic synthesis

5.1 Process identification and redesign

A perennial methodological flaw in automation projects is the unspoken assumption that the process you're observing is the process that should be. The mapping studies cited make it clear that the industrial landscape is highly heterogeneous and that technical viability is frequently assessed in isolation from organisational redesign [7]. For agentic applications this is particularly expensive. You need to be aware of the informal workarounds, the conflicting approval paths and the discretionary actions that are simply not in the formal documentation before you go to the trouble of modelling the process. An agent trained on the official process diagram is likely to get it completely wrong in exactly the circumstances where the human is actually the only thing that can do it. Process discovery therefore needs to be a combination of transaction logs, interviews and controlled observation and needs to be able to tell the difference between legitimate discretion and avoidable inconsistency.

The first requirement for intelligent automation is a suitable process rather than a sophisticated model. RPA scholarship has established that stable processes with accessible digital inputs and clear rules are the easiest to automate, whereas cross-system exceptions and poor quality documentation lead to rework [5–9]. Farinha and others have identified a number of suitability criteria, including feasibility, process description and the quality of the inputs and outputs [9]. The agentic extension shifts the question. The enterprise is no longer asking whether a step can be done by software, but rather whether an agent can legitimately interpret an ambiguous instruction, initiate a transaction and alter a durable record.

A useful categorisation would be deterministic execution, predictive decision support and bounded discretionary action. Deterministic flows such as the extraction of approved invoice fields are still very much in the realm of conventional scripts. Predictive decision support (e.g. forecasting demand, prioritising maintenance tickets) would be the sort of thing where you'd want to use a trained model

with a monitored error rate. Bounded discretionary action (e.g. negotiating a supplier exception, resolving an incomplete customer enquiry) may be the sort of thing where you'd allow an agent to take some action, but you'd still have to have a process contract in place that defines the inputs, the tools that are permissible, the output schema, the owner and the recovery route. Allowing an agent to take the initiative and do things in a way that isn't explicitly defined in the process contract is not an option.

This division is all about automation programmes which are principally driven by the estimated number of employee hours that will be saved. A high volume process may in fact be of very low strategic value if upstream errors keep cropping up; a low volume but high consequence process may be the one that warrants investment in traceability rather than in labour substitution. The studies on implementation success all point to the importance of project governance, standardisation and business engagement [8–10]. Process mining, event logs and exception taxonomy can all be used to inform the prioritisation of automation, but of course the data collection process itself has to be mindful of confidentiality and purpose limitation. The selection therefore needs to be a balance of feasibility, business importance, exposure to error and the degree of control that you have over the actions.

Table 1. Evidence streams and their contribution to the synthesis.

Evidence stream	Established contribution	Main interpretive boundary
Transformation theory [1–4]	Change, capabilities and human–AI tensions	Not direct evidence of agentic performance
Process automation [5–14,29,30]	Selection, governance, integration and value	Mostly rule-based or assisted automation
Agent architectures [15–20]	Planning, tools, memory and coordination	Primarily technical or conceptual evidence
Productivity studies [21,22]	Task-level or field productivity estimates	Assistive settings; limited autonomous action
Saudi studies [23–27]	Readiness, leadership and adoption conditions	Sparse multi-agent deployment observations
Language-model risks [28]	Bias, documentation and social consequences	General risks need context-specific testing

5.2 Agent orchestration and machine learning integration

The right level of decomposition also comes down to verifiability. A single agent with well defined, bounded tools is actually going to be easier to monitor than a network of conversational agents where you can't really trace the intermediate decisions. Multi-agent collaboration is only appropriate when the task genuinely requires different competencies, independent checking or parallel work. It shouldn't be used just because having a role based dialogue sounds cool. Agents need to have explicit termination conditions, limits on the number of times they can call a tool and safeguards against circular delegation.

Contemporary agent work is framed in terms of an execution cycle: take an objective, break it down into sub-tasks, use tools to complete those sub-tasks, observe the results and then revise the response as appropriate [15–19]. Multi-agent setups allow for specialisation and communication, with the potential to distribute the work of retrieval, validation and execution across a number of

different services [16,17]. This is very attractive in the context of enterprise workflows that span CRM, ERP and document repositories. The caveat is that conversation between agents is not an organisational control. Each message can contain an unvalidated assumption and the more agents you have the greater the potential for latency, cost and security exposure.

Machine learning should only be introduced into this architecture where it provides tangible predictive value. Classification can be used to identify the type of document, anomaly detection can be used to flag potentially suspicious transactions and predictive models can be used to estimate the urgency of maintenance. Language models can be used to interpret the request and orchestrate the appropriate tools; they should not be used to silently replace well-calibrated numerical models or hard rules. Reviews of agent behaviour reveal deficiencies in memory, generalisation and evaluation which become very important when actions are linked to enterprise systems [18,19]. A structured state representation and typed tool outputs go a long way to reducing the need for unstructured natural language interpretation.

A practical orchestration design is to separate out the planning service from the policy enforcement service and the execution connector. Prior to any tool call the policy service would check identity, purpose, transaction limits and approval status. Post execution the system would then check that the desired state change actually occurred. This also provides a clear answer as to what should happen when the model is uncertain: it should either request the missing evidence or escalate rather than make up a completion. These are the controls we are proposing as engineering requirements here; the literature we've reviewed provides no single 'best' architecture for Saudi enterprises.

5.3 Data architecture and interoperability

Enterprise wide automation is predicated on consistent identifiers, accessible metadata and a trusted history of transactions. A technically capable agent cannot reconcile conflicting customer records or authorise an action against data of unknown provenance. Research into enterprise systems and automation has shown that integration capability is far more important than the mere imitation of an interface when it comes to organisations wanting to achieve real and sustainable process improvement [6,8,30]. The data architecture therefore comes first. The key elements of such an architecture are authoritative system ownership, access to the documents, schema compatibility, retention policies and the maintenance of the business definitions in context.

Retrieval from enterprise knowledge sources must preserve the document version, source owner and access permissions. When an agent is drafting a procurement recommendation they need to be able to tell the difference between an approved contract clause and an obsolete template, and a supplier quotation and a financial system record. The same is true of Arabic and English documents where translation issues can impact the meaning of obligations, the meaning of measurements and the meaning of role titles. Data quality testing should look at completeness, timeliness, duplication and semantic consistency across the various departments. You shouldn't be able to infer model performance from a demonstration which is based on a set of very clean documents.

Interoperability also brings with it an architectural trade-off. A direct connection to a legacy user interface may get you to pilot much quicker but is going to be very fragile in the face of layout changes and hidden state. Stable interfaces, event driven integration and auditable service calls will generally give you much clearer control boundaries. RPA may still be justified in situations where the interfaces are not available, as long as you have monitoring in place to catch failures. The point is not to just go out and eliminate the old systems, it's to have a reliable record of the decisions and the changes that

are being made across the technology estate. Without that foundation you're just going to end up with a much faster replication of the organisational fragmentation.

5.4 Workforce transformation and human judgement

Adoption research into intelligent process automation indicates that trust and transparency are key factors in determining the willingness to adopt a system even if the technical merits are there [12]. Trust must not be equated with less scrutiny however. An employee who is continually presented with slick, polished but incorrect explanations may end up being completely untroubled by the outputs which are in fact professionally presented. Competence assessments should therefore look at challenge behaviour: do people go out and seek evidence? Do they spot the fake references? Do they spot the uncertainty? Do they record their disagreement? Managers need to protect the time for independent checking in the early days of deployment rather than simply demanding the same level of throughput that you would expect once the system has matured and is performing reliably. That is the sort of environment in which professional judgement is actually exercised rather than just going through the motions.

The evidence on AI-assisted work makes the simple 'displacement' story very difficult to sustain. The controlled studies show productivity gains on bounded writing tasks, the field evidence from customer service shows mixed results with gains that are clearly related to the experience of the worker and the nature of the task [21,22]. This is about assistance not about unbounded agent autonomy and so can't be directly generalised to Saudi manufacturing, banking or government decision making. It does however point to a mechanism: the system can provide useful information and reduce the need for the worker to go out and search for information. That in turn will change the mix of skills that is required to be a competent worker.

Human expertise is of critical importance at the exception boundaries. You need people to spot contradictory records, to question the adequacy of model explanations and to be able to stop things that have legal or financial implications. Raisch and Krakowski's paradox is that automation and augmentation should be co-designed [4]. If management are only measuring the reduction in processing time they may well be missing the increased review burden, the reduced opportunity to learn or the lack of accountability without authority. Organisational design therefore needs to specify who is allowed to approve, who is allowed to override, who is allowed to appeal and who is allowed to correct the automated outcome.

Saudi implementation adds in the form of different capacity needs. The large organisation may be in a position to set up their own specialist engineering and assurance teams whereas the small firm may be more inclined to go with a vendor and a shared support model. Research into the Saudi SMEs has shown that digital readiness and adoption are linked to leadership, infrastructure and awareness [23-25]. Translating that into the context of agentic AI is going to be a bit of a stretch as prompt literacy on its own is not enough. The people in the organisation will need to have the process knowledge, the data literacy, the cyber hygiene and the ability to spot when the model has failed. Effectiveness of the training will need to be measured in terms of the outcomes of simulated incidents and the actual decisions that are documented, not just in terms of attendance and self reported confidence.

5.5 Governance, security and institutional legitimacy

Operational assurance must make use of both preventive and detective controls. The preventive controls would be an approved tool manifest, credential isolation, validation of structured parameters, rate limiting and mandatory authorisation for any irreversible transaction. The detective controls would be anomaly alerts, post-execution reconciliation, adversarial testing and periodic review of the action logs. An agent which is able to accurately summarise a malicious instruction contained within a retrieved document may still be unsafe if it treats that instruction as of higher authority than the organisation's policy. The testing therefore needs to include prompt injection, conflicting evidence, unexpected tool errors and attempts to gain unnecessary privileges. Oversight should be judged on its ability to prevent or contain actual errors rather than on the mere existence of policy documents.

The move from advisory output to action creates a governance discontinuity. A model recommending that a payment be reviewed is a very different proposition to an agent actually going and making the payment via an authenticated account. The technical surveys all point to the dangers of tool misuse, coordination risks and unreliable outputs [15-20]. The critical literature on language models is equally clear on the risks of biased representation, opaque training data and uneven social costs [28]. All of which means that governance has to be a design input rather than a compliance activity which you tack on at the end of the deployment cycle.

An accountability model requires explicit decision rights, least privilege access, transaction ceilings, human approval thresholds and immutable event histories. Sensitive requests need to be checked against an authoritative record prior to execution and there should be a clear separation between the agent that makes the decision and the actor that authorises a high consequence change. Monitoring must capture attempted actions, rejected actions, source evidence, model/tool versions, reviewer identity and the final system state. This enables incident analysis without having to assume that the textual explanation that the model produces is a true reflection of its internal reasoning.

Governance must also be tuned to the organisational and sectoral context. The externalities of a logistics scheduling assistant are very different from those of a clinical decision support tool or a credit eligibility agent. Data access, retention and cross border processing will all need to be assessed in the light of the applicable organisational obligations and Saudi requirements; this review in no way suggests that one generic control will satisfy all the laws. The research into public sector digital transformation in Saudi Arabia points to the importance of management support and institutional arrangements, with organisational adoption evidence identifying a range of barriers which go well beyond the technology itself [26,27]. Legitimacy therefore will be contingent on contestability, service quality and clear accountability where automated action has the potential to disadvantage people.

5.6 Performance, value and scaling mechanisms

A better design would be to randomise the eligible cases between agent assistance and the existing workflow where that is possible. Where randomisation isn't an option you could consider a phased rollout, matched sites or an interrupted time series. The primary outcomes would be the validated task completion rate, end to end cycle time, consequential error rate and fully loaded cost per case resolved. The secondary outcomes would be appeal rate, employee learning, accessibility and recovery time after a failed action. You would need to be very clear about the denominator as

reducing the time taken to deal with the simple cases is of very limited value if the complex cases are simply being pushed into some unmeasured queue.

The claims around intelligent automation value are almost invariably focused on hours saved. Enterprise transformation is a much more complex affair and requires a much longer causal chain. The outcomes of interest are process reliability, correction cost, user experience, employee capability, auditability and cross-functional responsiveness. Research into RPA value creation has demonstrated that the operational impact varies by activity and is not uniform [11]. There is some evidence from controlled generative AI studies of real productivity gains in certain types of work [21,22], but again that doesn't necessarily point to agentic planning and doesn't establish any kind of sustained firm-wide benefit.

A defensible business case would involve comparing agent enabled workflows to some meaningful alternative (redesigned manual processes, conventional scripts, predictive decision support etc). The evaluation would need to take into account the cost of integration, the cost of model inference, the cost of human review, the cost of maintenance, the cost of failure recovery and the cost of security assurance. The benefits would have to be tied to actual business events rather than the number of pieces of text that were generated. Where the speed of service is improved but the amount of error remediation is increased, a single measure of productivity can be very misleading. It's absolutely essential that the baseline is comparable and that you have explicit exposure to changes in demand and changes in staffing.

Scaling introduces a second problem. The pilot may have succeeded because the expert was able to quietly fix things, go around missing integrations and curate the inputs. All of that supporting infrastructure tends to vanish when you scale out to other regions and other functions. Governance research into automation is full of references to the importance of ownership and coordination in moving beyond the local initiative [10,13,14,29]. Enterprise value therefore comes from institutionalising the reusable controls, the reliable interfaces and the feedback loops for improvement rather than simply having a number of independent assistants. The unit of analysis is the end to end service outcome, not the number of agents.

6 An accountable enterprise automation framework

Application domains are the perfect illustration of why uniform deployment policies are a non-starter. In procurement you might have an agent that's gathering up the supplier docs and making a recommendation as to what to do, with the actual procurement authority still lying with the appointed officers. In the industrial environment you could have a predictive model that's quite clearly going to tell you that a piece of equipment is deteriorating, but the decision as to whether you actually stop production would be one that would be governed by independent safety rules and the judgement of accountable engineering. In public administration you could have an agent that's doing document triage to speed up the service delivery process, with the actual eligibility decisions being subject to transparent rules and accessible review. These are design situations, not actual outcomes. The common thread is that the machine is capable of collating the evidence and the institution is the one that has the authority to make the actual decision.

Figure 1 shows a proposed layered architecture that links the authorised business demand to traceable execution. The intake layer identifies the request, the role and the business purpose. The context layer pulls the relevant permission-filtered records and establishes the provenance of the evidence. Planning selects from an approved tool catalogue to determine the sequence to be taken;

predictive services will provide a score or classification where applicable. A policy gate then assesses authority, risk, transaction thresholds and escalation conditions before any state changing action is permitted. Verified execution and event logging then feed into the exception management and continuous improvement loops. The policy gate is deliberately decoupled from the agent planner.

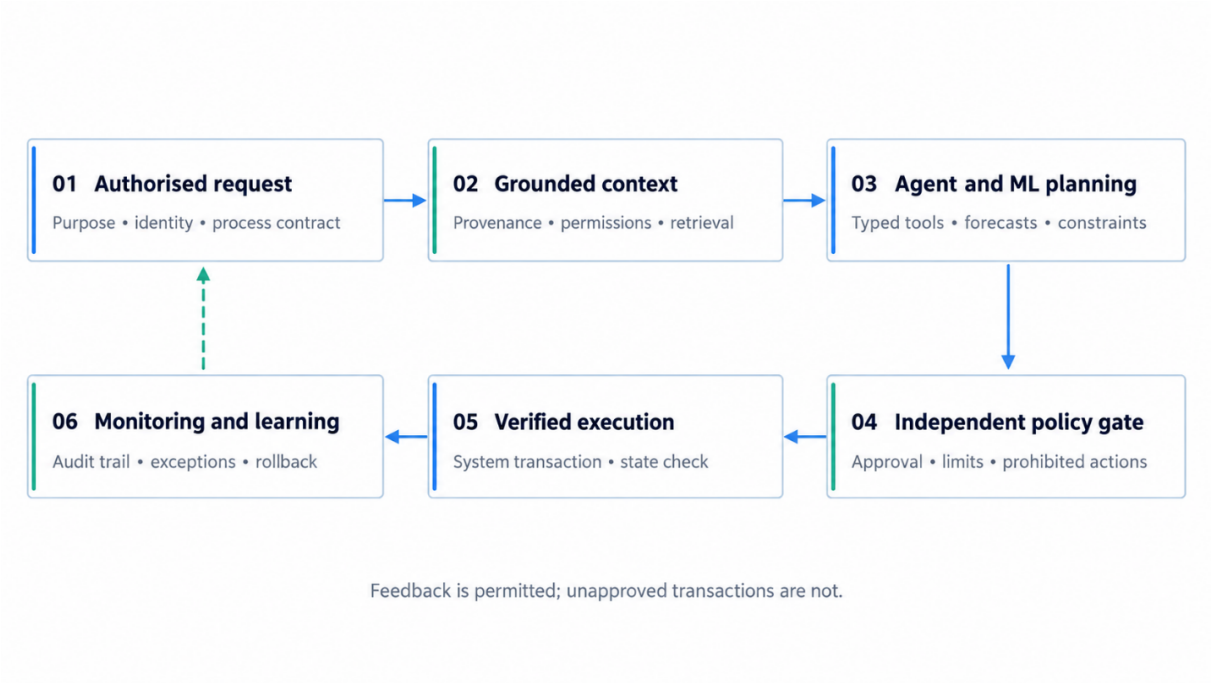


Figure 1. Proposed accountable enterprise automation architecture. Arrows indicate authorised execution and auditable feedback, not observed causal effects.

Autonomy is treated as a configurable property rather than a simple on/off switch. For low risk information tasks you can have automatic retrieval and summarisation. For moderate risk actions you can have some form of structured validation or indeed just sampled review. High risk transactions will always require named human approval and of course prohibited actions will simply not be available even if the model is capable of producing a very convincing justification. The ability to make this distinction enables organisations to go all out on speed where the risk is low and to maintain an appropriately accountable level of control over decisions of consequence. Figure 2 shows the path from foundational readiness to the point of monitored organisational outcomes.

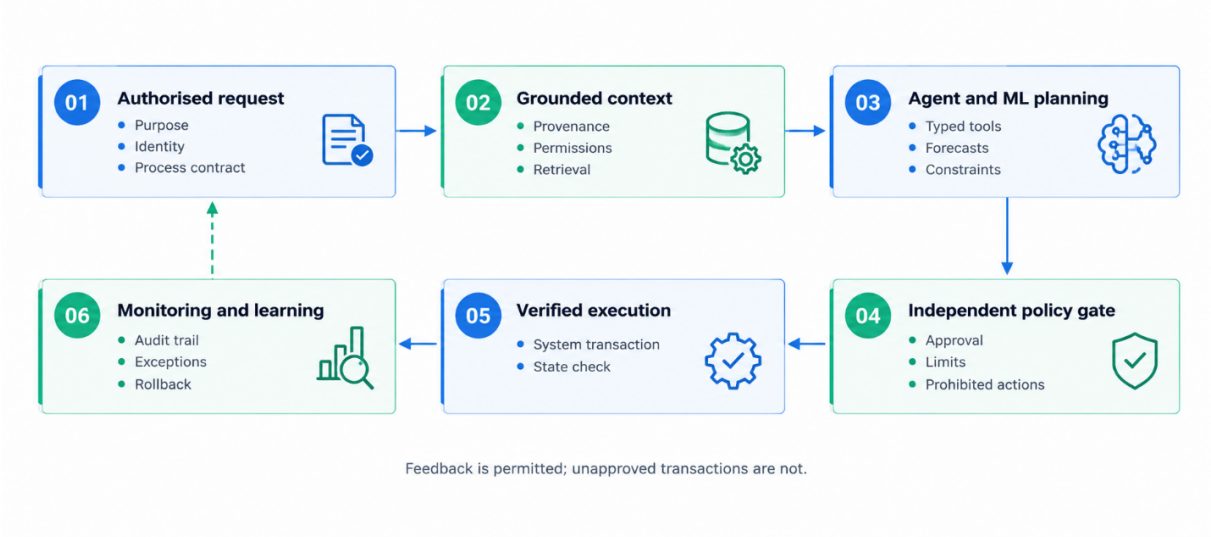


Figure 2. Proposed transition from foundational readiness to verified scale. Progress depends on independently assessed control and outcome evidence.

For Saudi organisations the application would start with a very narrow workflow contract indeed: defined inputs, authorised systems, affected stakeholders and an independent acceptance test. The contract would specify Arabic and English quality requirements where appropriate, evidence retention, escalation times and a mechanism for revoking tool access. It's a conceptual synthesis rather than a platform that has been deployed or experimentally validated. Its value is in making the assumptions and the responsibilities inspectable prior to any capital being committed.

7 Discussion and research gaps

The Saudi evidence base also cautions against the idea of treating firms as a homogenous group. The Ha'il evidence looks at the link between digital readiness and SME performance but its local and methodological constraints mean it's not going to be generalisable to all sectors [24]. The related manufacturing evidence looks at analytical readiness and decision making rather than at the actual execution of the tools [25]. A properly designed future study would look at the organisations in the sample by size, sector, degree of legacy system dependence and data governance maturity. That would tell you whether the poor results are a function of the agent technology itself or whether they are simply a result of the organisational conditions that prevent any digital intervention from working.

The evidence supports a conditional proposition: agentic systems can be a force for digital transformation where there is a match between technological discretion and organisational control. That's very different from the assertion that agentic systems will in themselves transform the enterprise. The agent surveys point to plausible capabilities, the RPA studies document the mechanisms of implementation and the Saudi work identifies the readiness constraints. None of those individually provide evidence of cross-sector causal gains from agentic AI as part of Vision 2030. The main synthesis is therefore a mechanism based explanation rather than a number.

Three research gaps are clearly of priority. First, longitudinal work is needed to see if bounded agents do better than optimised scripts on processes of varying exceptionality, with the same cost and error accounting. Second, we need to look at whether different human approval thresholds have an impact on harm, latency and trust (rather than just on satisfaction). Thirdly, multi-site work in Saudi Arabia

is required to see how the scale of the organisation, the presence of bilingual documentation, the level of integration with legacy systems and the degree of public/private interdependence all affect performance. Without those designs the claims about national transformation are still very much in the aspirational rather than the evidential.

Another gap is in the distribution of the benefits and the burdens. The studies should look at whether the productivity gains actually lead to improved access to services and improved employee development or whether they merely pass on the unmeasured review work to a smaller number of people. Security evaluations must include deliberately bad documents, tool permission changes and failed recovery scenarios. Publishing the negative results and having very clearly documented boundary failures would contribute far more to the cumulative knowledge than a number of studies which simply reiterate the successful demonstration of a given capability. The proposed framework provides the constructs to test these relationships without pre-determining what the outcomes will be.

Table 2. Operational evaluation and control matrix for staged deployment.

Dimension	Illustrative measure	Approval or review criterion
Process fidelity	Verified completion; documented exceptions	No unexplained state changes
Model performance	Task-appropriate error and calibration	Threshold agreed by process owner
Data assurance	Source freshness; provenance; access tests	Authorised and versioned evidence
Security	Blocked unsafe tool calls; incident recovery	Demonstrated containment and rollback
Human accountability	Escalation, override and appeal records	Named authority for consequential actions
Enterprise value	Cost per resolved case; time; rework	Net gain against credible baseline
Inclusion and skills	User accessibility; competency testing	No unmanaged burden transfer

8 Managerial and policy implications

A phased implementation programme can turn these principles into real decisions. Phase 1 is about defining the purpose of the process and establishing a benchmark. Phase 2 is about testing retrieval and recommendations without any permission to alter the records. Only once you've demonstrated error thresholds and recovery procedures should you go on to bounded execution. Each subsequent stage must be accompanied by a fresh assessment of security exposure, workload and the external stakeholders. That sequencing doesn't guarantee you'll succeed but it does create observable decision gates and enables you to kill an underperforming investment before you've become inextricably tied up in the dependencies.

Enterprise leaders should be taking a portfolio approach to candidate evaluation, scoring each one on process stability, strategic relevance, data quality, level of autonomy required and the severity of the consequences. The governance board will need representation from ops, technology, legal assurance and the business functions that are actually affected by the process. Investment decisions will need to be based on baselines, rollback mechanisms and evidence that the agentic design actually delivers value over a simpler alternative. The procurement terms need to cover data handling, service continuity, auditing rights and the portability of the process records.

For policymakers and the various ecosystem institutions the priority is to provide reliable foundations rather than to prescribe a particular model or vendor. Shared capability standards,

interoperable records, applied training and support for SMES can go a long way to reduce the disparity in readiness. Public sector procurement can be used to drive transparent assessment and human contestability. Alignment with Vision 2030 should be demonstrated by demonstrable outcomes in terms of productivity, service and capability rather than simply invoked as a get out of jail free card.

9 Limitations

This review makes use of a deliberately selected and DOI verified literature set rather than a full search so there is a potential for some of the less visible studies and non-English work to be missed. A large proportion of the agent research is conceptual or benchmark type and the Saudi empirical work is largely concerned with digital maturity or general AI adoption. As such some of the architectural recommendations go beyond the direct evidence of what has been tested. The framework has not been prospectively field tested, risk quantified or subjected to formal expert consensus. The recommendations are therefore to be adapted to the specific obligations of the industry and the capabilities of the organisation.

10 Conclusion

Agentic AI opens up new possibilities for coordinating complex work, but genuine enterprise transformation will only be possible if all of the right conditions are in place. The evidence reviewed points to a set of interdependent conditions: appropriate processes, reliable data, bounded execution, trained people, accountable governance and outcome based evaluation. The proposed framework ties these conditions together in a controlled sequence from authorised request through to action and organisational learning. In the Saudi context this provides a route to link enterprise innovation with Vision 2030 without making any premature claims about productivity or national impact. The next step is to test the framework in real organisations, compare it to simpler forms of automation and be transparent about both the successes and the failures. Responsible scale is going to be determined by what an organisation can verify, correct and sustain rather than simply what the agents can attempt.

References

[1] Verhoef PC, Broekhuizen T, Bart Y, et al. Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*. 2021;122:889–901. https://doi.org/10.1016/j.jbusres.2019.09.022

[2] Hanelt A, Bohnsack R, Marz D, Antunes Marante C. A systematic review of the literature on digital transformation: Insights and implications for strategy and organizational change. *Journal of Management Studies*. 2021;58:1159–1197. https://doi.org/10.1111/joms.12639

[3] Kraus S, Durst S, Ferreira JJ, et al. Digital transformation in business and management research: An overview of the current status quo. *International Journal of Information Management*. 2022;63:102466. https://doi.org/10.1016/j.ijinfomgt.2021.102466

[4] Raisch S, Krakowski S. Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*. 2021;46:192–210. https://doi.org/10.5465/amr.2018.0072

[5] Hofmann P, Samp C, Urbach N. Robotic process automation. *Electronic Markets*. 2020;30:99–106. https://doi.org/10.1007/s12525-019-00365-8

[6] Syed R, Suriadi S, Adams M, et al. Robotic process automation: Contemporary themes and challenges. *Computers in Industry*. 2020;115:103162. https://doi.org/10.1016/j.compind.2019.103162

[7] Enríquez JG, Jiménez-Ramírez A, Domínguez-Mayo FJ, García-García JA. Robotic process automation: A scientific and industrial systematic mapping study. *IEEE Access*. 2020;8:39113–39129. https://doi.org/10.1109/ACCESS.2020.2974934

[8] Herm LV, Janiesch C, Helm A, et al. A framework for implementing robotic process automation projects. *Information Systems and e-Business Management*. 2023;21:1–35. https://doi.org/10.1007/s10257-022-00553-8

[9] Farinha D, Pereira R, Almeida R. A framework to support robotic process automation. *Journal of Information Technology*. 2024;39:149–166. https://doi.org/10.1177/02683962231165066

[10] Schlegel D, Rosenberg B, Fundanovic O, Kraus P. How to conduct successful business process automation projects? An analysis of key factors in the context of robotic process automation. *Business Process Management Journal*. 2024;30:99–119. https://doi.org/10.1108/BPMJ-06-2023-0465

[11] Durão D, dos Reis AP. How does robotic process automation create value for firms? *Information Systems and e-Business Management*. 2024;22:721–740. https://doi.org/10.1007/s10257-024-00685-z

[12] Mayr A, Stahmann P, Nebel M, Janiesch C. Still doing it yourself? Investigating determinants for the adoption of intelligent process automation. *Electronic Markets*. 2024;34:56. https://doi.org/10.1007/s12525-024-00737-9

[13] Lacity M, Willcocks L, Gozman D. Influencing information systems practice: The action principles approach applied to robotic process and cognitive automation. *Journal of Information Technology*. 2021;36:216–240. https://doi.org/10.1177/0268396221990778

[14] Cascais Brás J, Pereira RF, Melo M, Bianchi IS, Ribeiro R. Balancing business, IT, and human capital: RPA integration and governance dynamics. *Information*. 2025;16:793. https://doi.org/10.3390/info16090793

[15] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*. 2024;18:186345. https://doi.org/10.1007/s11704-024-40231-1

[16] Li X, Wang S, Zeng S, Wu Y, Yang Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*. 2024;1:9. https://doi.org/10.1007/s44336-024-00009-2

[17] Guo T, Chen X, Wang Y, et al. Large language model based multi-agents: A survey of progress and challenges. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*; 2024. p. 8048–8057. https://doi.org/10.24963/ijcai.2024/890

[18] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*. 2025;68:121101. https://doi.org/10.1007/s11432-024-4222-0

[19] Plaat A, van Duijn MJ, van Stein N, et al. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*. 2025;84:29. https://doi.org/10.1613/jair.1.18675

[20] Gao C, Lan X, Li N, et al. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*. 2024;11:1259. https://doi.org/10.1057/s41599-024-03611-3

[21] Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 2023;381:187–192. https://doi.org/10.1126/science.adh2586

[22] Brynjolfsson E, Li D, Raymond L. Generative AI at work. *Quarterly Journal of Economics*. 2025;140:889–942. https://doi.org/10.1093/qje/qjae044

[23] Badghish S, Soomro YA. Artificial intelligence adoption by SMEs to achieve sustainable business performance: Application of technology–organization–environment framework. *Sustainability*. 2024;16:1864. https://doi.org/10.3390/su16051864

[24] Tripathi A, Singh A. SMEs awareness and preparation for digital transformation: Exploring business opportunities for entrepreneurs in Saudi Arabia's Ha'il region. *Sustainability*. 2024;16:3831. https://doi.org/10.3390/su16093831

[25] Aldossari S, Mokhtar UA, Abdul Ghani AT. Empowering Saudi manufacturing small and medium enterprises: A framework for big data analytics adoption and its impact on decision-making. *SAGE Open*. 2025;15. https://doi.org/10.1177/21582440251369162

[26] Saqqat AA. Digital transformation in public organizations: An applied study on the Ministry of Health in the Kingdom of Saudi Arabia. *Journal of Economic, Administrative and Legal Sciences*. 2024;8:120–137. https://doi.org/10.26389/AJSRP.Q040623

[27] Saeede AY. Trends in the adoption of artificial intelligence in knowledge management: An analytical study using the technology acceptance model in Saudi organizations in line with Vision 2030. *Journal of Economic, Administrative and Legal Sciences*. 2025;9:69–84. https://doi.org/10.26389/AJSRP.R170225

[28] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*; 2021. p. 610–623. https://doi.org/10.1145/3442188.3445922

[29] Durão D, dos Reis AP. An organizational perspective on robotic process automation adoption and usage factors. *Applied System Innovation*. 2025;8:33. https://doi.org/10.3390/asi8020033

[30] Dumitru VF, Ionescu BȘ, Rîndașu SM, Barna LEL, Crîjman AM. Implications for sustainability accounting and reporting in the context of the automation-driven evolution of ERP systems. *Electronics*. 2023;12:1819. https://doi.org/10.3390/electronics12081819