The Agentic Workforce Operating Model: Redesigning Decision Rights, Controls and Productivity

A Board Framework for Accountable Human-Agent Work

Authored by Chennakeshav Adya
Independent Researcher
September 2026

Abstract

Agentic artificial intelligence extends workplace automation from generating content or recommendations to planning steps, selecting tools and taking actions across digital systems. That change can shorten cycle times and release capacity, but it also changes who may decide, who may commit the organisation, how evidence is retained and how failures are contained. Conventional process maps and access-control models often assume that a human initiates each material action. An agentic operating model needs a more explicit allocation of authority.

This paper develops a practical framework for boards, executives and transformation leaders designing human-agent work. It separates five forms of authority: observe, recommend, prepare, execute and commit. It then links authority to consequence, reversibility, evidence quality, system access and the organisation's ability to detect and contain error. The model combines decision-rights design, control tiers, identity and access management, human oversight, process redesign, workforce adoption and a productivity scorecard.

A hypothetical shared-services programme illustrates the framework. The author assumes a business processing 480,000 cases a year across finance, procurement, customer operations and workforce administration. All volumes, costs, time savings and implementation outcomes in the example are analytical assumptions. They are neither observed company data nor forecasts. The example shows why activity automation does not automatically create financial value: released time must translate into lower external spend, avoided hiring, increased throughput, better service or redeployed capacity. The central conclusion is that agentic productivity depends on disciplined authority and process design. Organisations should scale autonomy only when they can identify the accountable owner, reconstruct the agent's actions, enforce limits and demonstrate a durable outcome.

JEL Classification: D23, J24, L22, M12, M15, O32, O33

Keywords: agentic AI, operating model, decision rights, human oversight, internal controls, productivity, workforce redesign, AI governance, process transformation

1. Start with the operating decision

The board's first decision is where agentic systems may exercise operational authority and under what conditions. This is broader than choosing a model or approving a technology budget. It determines how work will move, how customers and employees may be affected, which systems an agent may reach, who remains accountable and how a material action can be stopped or reversed.

An AI assistant produces material for a person to assess. An agentic system can interpret a goal, plan intermediate steps, retrieve information, use software tools and change a business record. The distinction is practical rather than promotional. A workflow becomes more agentic as the system gains discretion over sequence, tools, counterparties, timing or execution. Each additional degree of discretion changes the control requirement.

The starting question should be expressed as a business decision. Examples include whether to automate invoice exception handling, prepare supplier negotiations, resolve low-risk customer requests, reconcile cash, monitor covenants or assemble a board pack. The sponsor should define the outcome, present baseline, affected stakeholders, acceptable error, maximum exposure, required evidence and economic objective. A broad ambition to deploy agents across the enterprise is not an operating case.

NIST's AI Risk Management Framework asks organisations to define the tasks an AI system will support, the context of use, human roles and oversight, and the risks and benefits across the system lifecycle [1]. Its Generative AI Profile extends that approach to risks associated with generative systems and complex value chains [2]. These frameworks support a disciplined opening: map the work and its consequences before choosing autonomy.

The decision also needs an alternative. A process may improve through standardisation, simpler policy, conventional automation, better data, additional training, outsourcing or removal of a redundant step. Agentic technology should be selected when its ability to interpret variable inputs and coordinate actions creates an advantage that exceeds the added control, evaluation and operating cost.

The output is an approved use-case charter. It names the accountable executive, process owner, system owner, risk owner, affected population, authority boundary, success measures, pilot perimeter, funding and stop conditions. This charter becomes the reference point when scope expands or the system's capabilities change.

2. Define agentic work as a chain of delegated authority

An organisation delegates work through roles, policies, systems and approval limits. Agentic systems should enter that structure through a documented mandate. The mandate describes what the agent may do, where it may act and when a person must intervene. It should be narrow enough to enforce and broad enough to support the intended outcome.

Five authority levels provide a useful grammar. Observe allows the system to read approved information and detect conditions. Recommend permits it to propose a decision with evidence. Prepare permits it to draft records, messages or transactions without releasing them. Execute permits a reversible action within a defined limit. Commit permits the organisation to an external, legal, financial, employment or safety consequence. The last level requires the strongest case because recovery may be costly or impossible.

Authority is multi-dimensional. An agent permitted to create a purchase order up to USD 5,000 may still be restricted to approved categories, suppliers, cost centres, jurisdictions and operating hours. It may need verified supporting documents and an absence of related-party or sanctions flags. A nominal value limit alone does not describe its mandate.

The mandate should cover purpose, inputs, tools, data classes, systems, action types, counterparties, monetary limits, frequency, duration, geography and escalation triggers. It should also specify forbidden actions. This converts policy into a testable control object. The organisation can compare the intended mandate with configured permissions, observed tool use and actual outcomes.

NIST's core framework calls for policies that differentiate roles and responsibilities in human-AI configurations and for executive leadership to take responsibility for AI risk decisions [3]. The EU AI Act requires effective human oversight for high-risk systems and states that assigned overseers need competence, training, authority and support [4]. These sources apply in defined contexts, yet the operating principle travels well: oversight is credible when the person can understand, challenge, override and stop the system.

Figure 1. Human-agent decision rights progress from observation to organisational commitment

Authority expands only when evidence, reversibility, monitoring and accountable ownership support the consequence.

Table 1. Decision-rights grammar for agentic work

Authority	Permitted activity	Required evidence	Human role
Observe	Read approved sources, detect and classify	Source, timestamp, identity and confidence	Review material exceptions
Recommend	Propose an action and explain the basis	Inputs, rule, alternatives and uncertainty	Decide or request more evidence
Prepare	Draft a record, transaction or communication	Complete audit trail and validation checks	Approve release
Execute	Perform a bounded, reversible action	Precondition tests, limit check and confirmation	Monitor and manage exceptions

Authority	Permitted activity	Required evidence	Human role
Commit	Create external or difficult-to-reverse consequence	Independent verification, explicit authority and retained evidence	Approve or retain direct control

The examples organise operating-model design. Applicable legal and regulatory requirements require case-specific advice.

3. Classify consequences before allocating autonomy

Autonomy should follow consequence rather than technical capability. A system may be able to send a message, approve a refund, change a price or terminate access. The operating model asks whether it should do so in a particular context and what loss could result.

Consequence can be assessed across financial loss, legal commitment, regulatory breach, privacy, security, safety, employment, customer treatment, reputation and operational continuity. The analysis should include direct and second-order effects. A small erroneous payment may reveal a control weakness that can be repeated at scale. An inaccurate customer response may create a regulated representation. A scheduling decision may affect worker health or protected groups.

Reversibility changes the response. A draft can be discarded. A journal entry may be reversed but still affect reporting and audit. A public message, disclosure of confidential data, safety action or employment decision may not be fully recoverable. Time to detect and time to contain matter as much as nominal value. A low-value action repeated thousands of times before detection can create a material exposure.

The organisation should define a consequence class for each action, not for the agent as a whole. The same system may observe unrestricted operational data, prepare a limited transaction and be prohibited from committing it. Tool-level and action-level controls provide a more accurate boundary than a single label such as low-risk agent.

The European Union framework is risk-based and imposes specific obligations on systems classified as high risk [4]. Data-protection law can also constrain solely automated decisions that produce legal or similarly significant effects, subject to the applicable regime and facts [5]. Employment, consumer, financial-services and safety rules may create further requirements. The operating model should maintain a jurisdiction and use-case obligations register rather than assuming that one global classification resolves every deployment.

An impact assessment should record affected groups, intended benefit, foreseeable misuse, failure modes, evidence, controls and residual risk. The UK AI Management Essentials tool asks organisations to maintain an AI system record, identify roles, assess impact, manage data, test systems and monitor performance [6]. This provides a useful minimum structure even where the tool is voluntary.

4. Build control tiers that match autonomy and reversibility

A control tier combines the action's consequence with the quality of the environment in which it operates. A stable, structured process with reliable data and deterministic validation can support more autonomy than a novel process with ambiguous policy and unstructured inputs. Control design should therefore consider both what can go wrong and how confidently the organisation can detect it.

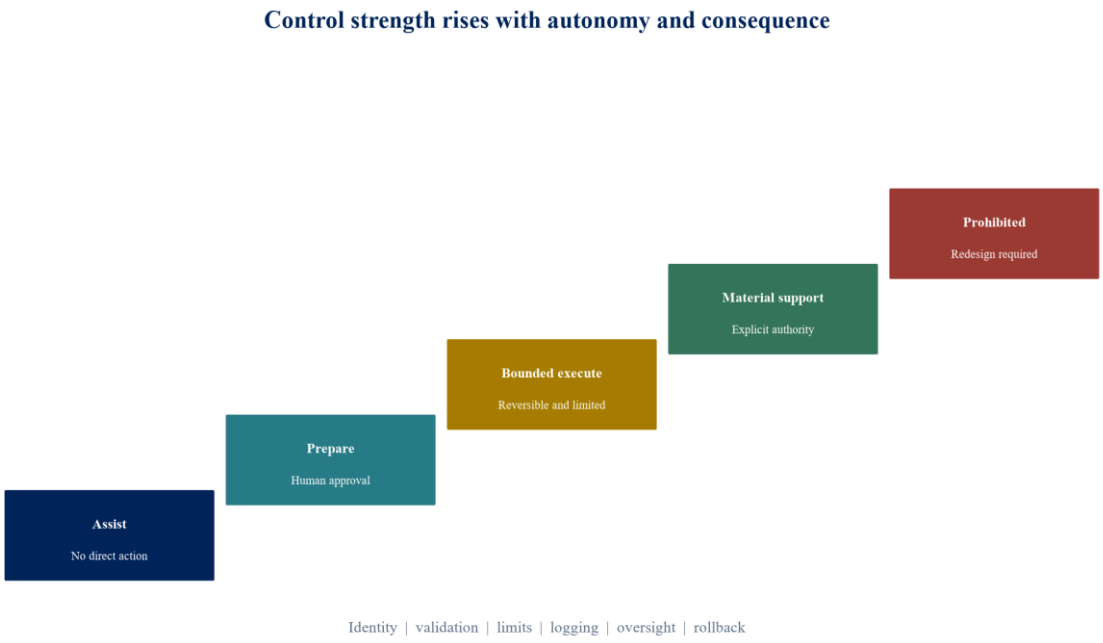
Tier 0 covers assistance without system action. The user remains responsible for assessing output. Tier 1 covers prepared actions that enter a human approval queue. Tier 2 covers bounded execution where preconditions are machine-verifiable, the action is reversible and monitoring is immediate. Tier 3 covers material or difficult-to-reverse actions and retains explicit human approval. Tier 4 covers prohibited autonomy because the organisation lacks evidence, control or acceptable risk capacity.

Each tier should specify mandatory controls. These can include approved data sources, least-privilege access, separation of duties, transaction limits, dual approval, deterministic validation, restricted counterparties, sandboxing, logging, anomaly detection, sampled review, rate limits, rollback, incident response and periodic re-authorisation. Controls should be tested as a system. A human approval step adds little when the reviewer sees no evidence, lacks time or routinely accepts the recommendation.

The UK National Cyber Security Centre advises organisations adopting agentic AI to assess the required autonomy, grant limited access, maintain visibility and preserve meaningful human control [7]. Its secure-development guidance also emphasises access controls, threat modelling, monitoring, incident procedures and secure defaults [8]. NIST's research on agent hijacking shows why untrusted data and trusted instructions require separation when an agent can use tools [9]. These sources make the control case concrete: permissions and observability determine the blast radius of a manipulated or mistaken action.

Control tiers should be re-evaluated after material changes to the model, prompts, tools, data, supplier, process, jurisdiction or user population. A pilot result applies to the tested configuration. It does not provide permanent authority for an evolving system.

Figure 2. The control staircase links autonomy to consequence and reversibility



Higher autonomy requires stronger identity, validation, monitoring, evidence and recovery capability.

Table 2. Control tiers for agentic execution

Tier	Operating permission	Conditions	Core controls
0	Generate or analyse for human use	No direct business-system action	Source visibility, user training and acceptable-use policy
1	Prepare action for approval	Reviewer can assess evidence before release	Approval queue, separation of duties and retained rationale
2	Execute bounded reversible action	Preconditions are verifiable; exposure is limited	Least privilege, limits, allowlists, logging, rollback and sampling
3	Support material action with explicit approval	Consequence is material or recovery is difficult	Independent evidence, named approver, dual control and incident readiness
4	No autonomous action	Evidence, control or risk capacity is inadequate	Sandbox, simulation or prohibition

Limits are approved for individual actions and use cases rather than assigned to a technology label.

5. Treat identity and access as the constitutional layer

Every agent that acts should have a unique identity. Shared service accounts obscure accountability and make revocation difficult. The identity should link to an owner, purpose, approved tools, data access, action limits, environment, validity period and current version. Credentials should be short-lived where feasible, stored securely and unavailable to the model's natural-language context.

Least privilege requires more than giving an agent the same access as its human sponsor. A person may have broad access because judgment, training and employment obligations constrain use. An agent can process instructions at machine speed and may be manipulated by content it reads. Its permissions should correspond to the narrow workflow and action set. Separate identities may be needed for reading, preparing and executing.

Tool interfaces should enforce typed, validated actions. An instruction such as manage this account is too broad. A safer interface exposes specific functions such as retrieve balance, prepare transfer within an approved beneficiary list or open an exception case. The tool should independently validate limits and required fields. Natural-language policy inside a prompt cannot replace an external enforcement point.

The operating environment should separate trusted instructions from untrusted content. Email, documents, websites and customer messages can contain malicious or accidental instructions. An agent should treat them as data unless a controlled channel explicitly grants authority. High-impact tools should not accept unverified parameters extracted from untrusted text without validation.

Access reviews should examine dormant agents, unused permissions, failed actions, unusual tool sequences, changes in volume and ownership departures. Revocation needs an immediate path. A central inventory should show every production agent and the systems it can affect. The inventory supports incident response, audit, supplier management and board reporting.

6. Make human oversight an operating capacity

Human oversight is a designed role with workload, information, competence and authority. It should not be added as a label after a workflow is built. The organisation must estimate how many cases require review, how long a sound review takes and whether the reviewer can identify a plausible error.

Automation bias arises when people over-rely on system output, especially under time pressure or when the explanation appears confident. Effective review should present the source evidence, important assumptions, uncertainty, policy rule, exception and proposed action in a form that supports challenge. Reviewers should be able to request additional evidence, change the result, reject the action and stop the workflow.

The EU AI Act's human-oversight provisions explicitly address understanding limitations, detecting anomalies, avoiding over-reliance, interpreting output and disregarding or reversing it where appropriate [4]. NIST similarly highlights the need to define and document human-AI roles and oversight [3]. These are useful design tests across many contexts, even when a particular legal provision does not apply.

Oversight can occur before action, during execution or after action. Pre-action approval suits material commitments. Concurrent supervision suits dynamic processes where intervention must be immediate. Post-action sampling can suit high-volume, low-consequence and reversible work when automated controls are strong. A blended design may route routine cases to bounded execution and material exceptions to prior approval.

The reviewer must have a stopping mechanism that works. This includes pausing one case, disabling a tool, revoking the agent's identity and stopping the service. The incident plan should specify who can use each mechanism, how continuity is maintained and how affected records are identified. A theoretical override that cannot be exercised quickly is not meaningful control.

7. Preserve evidence sufficient to reconstruct action

Agentic work produces a chain of decisions. The organisation should be able to reconstruct the goal, authorised user, agent version, source data, retrieved material, tool calls, validation results, approvals, outputs, business-system changes, exceptions and final outcome. This is operational evidence, not a verbatim transcript retained without purpose.

Logging should be designed around accountability and privacy. Sensitive prompts and data may require minimisation, masking, restricted access and retention limits. The organisation should decide which evidence is needed to investigate an event, demonstrate control, reproduce a result or satisfy audit and legal requirements. It should protect logs against tampering and connect them to existing security and business records.

Observability should detect behaviour that matters. Examples include attempts to use an unapproved tool, repeated validation failures, unusual action volume, new counterparties, escalating spend, high override rates, unexpected data access, deteriorating quality or excessive human review. Thresholds should reflect both individual and cumulative exposure.

Evidence also supports learning. A structured exception record shows which policy is unclear, which data is missing, where the model fails and which cases should remain human-led. The organisation can improve the process without concealing adverse outcomes. Control owners

should review false positives as well as missed risks because excessive blocking can destroy adoption and productivity.

The evidence architecture should remain usable when suppliers change. The organisation needs access to its own operational records and should understand what a vendor logs, retains and can provide after termination. Contract, architecture and continuity planning should address this before deployment.

8. Redesign the complete workflow

Adding an agent to one activity may accelerate that activity and leave the overall process unchanged. Productivity emerges when hand-offs, queues, approvals, rework and exception handling are redesigned around the outcome. The process owner should map the current state using actual volume and timing data, then identify which constraints determine end-to-end performance.

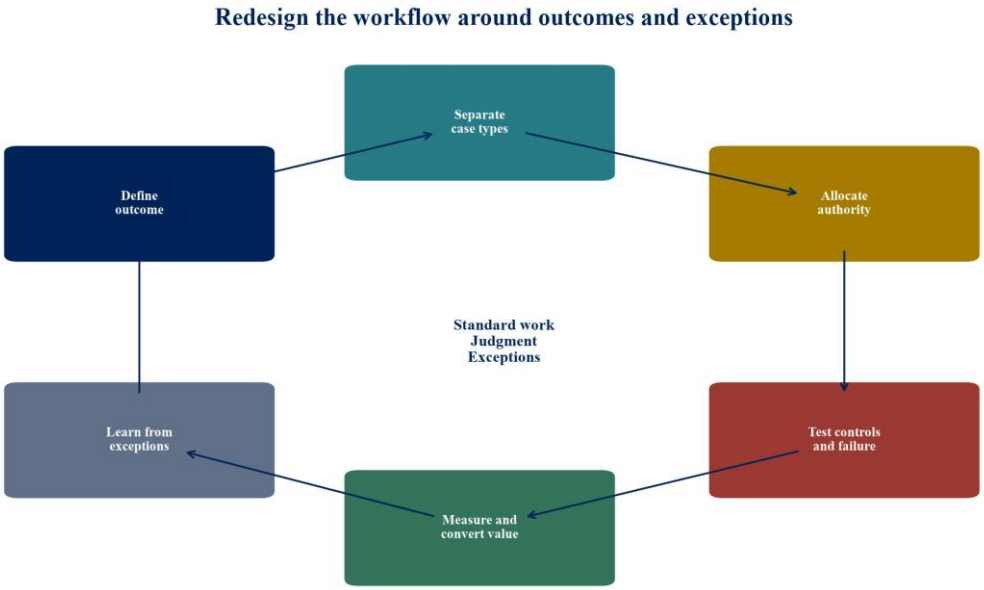
The redesign should separate standard cases, judgment cases and exceptions. Standard cases have clear policy, reliable inputs and testable outputs. They are candidates for conventional automation or bounded agentic execution. Judgment cases require interpretation and may benefit from evidence assembly and recommendation. Exceptions expose missing information, conflicting rules or unusual consequences and should route to an accountable person.

The future-state map should show which actor performs each step, what information is required, what decision is made, which system changes, what evidence is retained and where the case can stop. It should remove duplicate checks created by mistrust. It should also preserve a necessary independent control when the same agent creates and validates an action.

Process redesign often reveals policy debt. Employees may rely on tacit rules, local spreadsheets or informal escalation. An agent cannot safely execute ambiguity at scale. The sponsor may need to simplify products, clarify thresholds, clean master data, standardise contracts or assign policy ownership before increasing autonomy.

The UK government's human-centred guidance for scaling AI stresses integration into daily routines, training, risk management and monitoring [10]. OECD research on AI adoption reports that managers can struggle to connect AI to real workplace problems and underestimate enterprise-wide organisational change [11]. These findings support a joint technology-and-operating-model programme rather than an isolated tool rollout.

Figure 3. Agentic process redesign begins with the outcome and closes through evidence and exceptions



Standard work, judgment and exceptions receive different authority and control paths.

9. Measure productivity as an economic bridge

Productivity should be measured from inputs to outcomes. Model accuracy, task completion and user adoption are operating indicators. Financial value requires a bridge from those indicators to capacity, cost, revenue, working capital, loss or risk. The finance team should agree the bridge before the pilot.

The baseline should include volume, demand mix, full-time equivalent effort, external spend, cycle time, error, rework, service level, control exceptions and outcome quality. A before-and-after comparison can be misleading when case mix, staffing or demand changes. Where feasible, phased rollout, matched groups or another credible comparison should isolate the effect.

The NBER study Generative AI at Work examined a staggered rollout to 5,179 customer-support agents and reported a 14 percent average increase in issues resolved per hour, with larger gains among novice and lower-skilled workers [12]. The result concerns a specific assistant, workforce and outcome. It demonstrates both the possibility of measured gain and the importance of heterogeneity. It does not establish an expected return for an autonomous workflow.

Released hours are not cash. Capacity becomes value when the organisation avoids hiring, reduces contractor spend, increases completed volume, improves service or redeploys people to an activity with measured benefit. Finance should record the conversion mechanism, owner and timing. It should also include model fees, integration, data work, monitoring, review, security, change management and incident cost.

Quality and control can improve value even when headcount remains stable. Faster resolution may increase customer retention. Better evidence may reduce loss or audit effort. More consistent

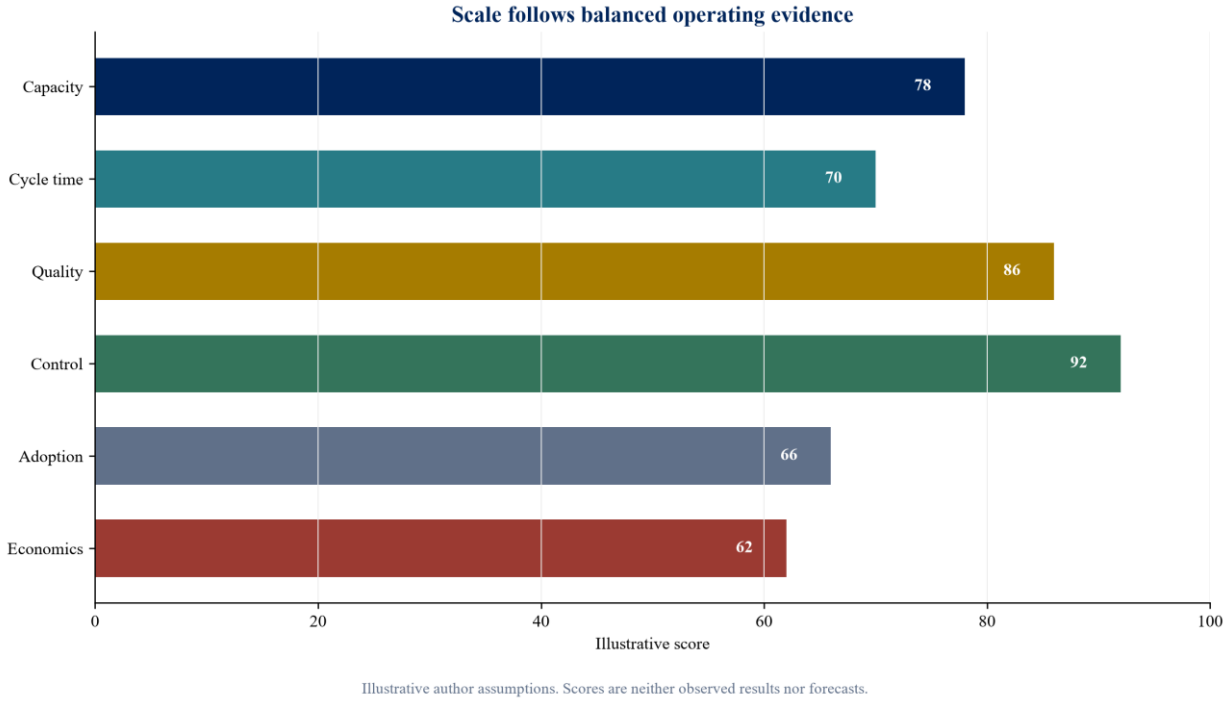
processing may support growth without proportional staffing. These effects need defined measures and a credible attribution method.

Table 3. Agentic productivity scorecard

Dimension	Baseline question	Operating measure	Financial bridge
Capacity	How much demand and effort exist by case type?	Cases per paid hour and queue age	Avoided hiring, reduced spend or added throughput
Cycle time	Where does work wait?	End-to-end elapsed time and service attainment	Working-capital, service or revenue effect
Quality	What constitutes a correct outcome?	First-time-right, rework, complaints and downstream defects	Loss avoided, retention or cost reduction
Control	Which failures matter?	Overrides, exceptions, breaches and detection time	Expected loss and assurance cost
Adoption	Who uses the workflow effectively?	Eligible usage, completion and escalation by cohort	Realised versus available benefit
Economics	What is the fully loaded cost?	Model, cloud, people, tooling and support per outcome	Contribution and payback

The scorecard combines output, quality, control, people and financial conversion.

Figure 4. Productivity is realised only when operating gains convert into durable outcomes



Capacity, cycle time and quality are tested alongside control, adoption and fully loaded economics.

10. Design the workforce transition around tasks and accountability

Workforce planning should begin at task level. The International Labour Organization's 2025 global index concludes that occupational transformation is more likely than full automation for many exposed jobs because occupations contain varied tasks and human involvement remains

important [13]. This points toward role redesign, skills and transition planning rather than a simple count of jobs removed.

Each affected role should be decomposed into tasks that remain human-led, become AI-assisted, move to agent preparation, qualify for bounded execution or disappear after process simplification. The analysis should identify new work in exception management, evaluation, policy ownership, system supervision and improvement. It should also identify skills that may erode if people no longer practise foundational tasks.

Role descriptions and performance measures need revision. A reviewer accountable for high-impact exceptions should be recognised for judgment and control, not only throughput. A process owner should own the combined human-agent outcome. Technical teams should not become the default owners of business decisions merely because they configure the system.

Training should be role-specific. Users need to understand capability, limitations, data rules and escalation. Reviewers need to interpret evidence, detect failure and resist automation bias. Control teams need access to logs and test results. Executives need to understand the operating and financial assumptions behind the programme.

Worker consultation can improve design by revealing tacit work, hidden exceptions and service consequences. OECD research based on more than 6,000 firms across six countries found widespread algorithmic-management use and reported managerial concerns including unclear accountability and limited explainability [14]. The same research records firm-level measures such as guidelines, audits, impact assessments, worker consultation and complaint channels. Consultation should be meaningful for the affected workforce and comply with applicable employment arrangements.

11. Govern third-party agents and model supply chains

Many agentic systems combine a foundation model, orchestration framework, cloud service, connectors, business applications and implementation support. The enterprise remains accountable for the deployment it chooses and should understand how each supplier affects security, continuity, evidence, cost and change control.

Supplier diligence should cover model and service versions, data use, retention, training, geographic processing, subprocessors, identity architecture, encryption, incident notification, evaluation, vulnerability handling, service levels, capacity, termination, portability and audit evidence. Contractual promises should align with the technical configuration. A no-training term has limited value when a connector sends data through a different route.

Change management is especially important. A supplier may update the model or safety policy, deprecate a tool, change latency or price, or alter geographic availability. The organisation should know which changes trigger re-evaluation, user communication or suspension. Material workflows may require version pinning, acceptance testing or a tested alternative.

Concentration should be assessed across models, cloud, identity, data and applications. A fallback shown in architecture should be tested with a representative workload. Switching may change output quality, tool behaviour, evaluation thresholds, regulatory evidence and cost. The continuity plan should estimate recovery time and manual capacity.

ISO/IEC 42001 specifies requirements for an AI management system and supports a lifecycle approach to roles, risk, objectives and continual improvement [15]. ISO certification or supplier alignment can provide useful evidence, while it does not replace evaluation of the actual use case, configuration and contractual boundary.

12. Establish clear operating ownership

The operating model needs named accountability from board oversight to case execution. The board approves risk appetite and material deployments. An executive sponsor owns business outcomes and resources. The process owner owns end-to-end performance and policy. The system owner controls configuration, release and reliability. Risk, legal, privacy, security and internal control functions provide challenge and requirements within their mandates. Internal audit may assess whether governance and controls operate as described.

A central AI function can maintain standards, inventory, evaluation methods, approved platforms and common controls. It should not absorb accountability for every business outcome. Federated process owners need enough capability to operate within the framework. The balance depends on organisational scale, regulation, technical maturity and use-case diversity.

The responsibility matrix should distinguish ownership, approval, operation, review and assurance. Multiple people can contribute, but one role should own each material decision. The matrix should include incident classification, emergency shutdown, customer remediation, regulatory notification, model change and retirement.

Management information should report a portfolio rather than isolated pilots. Useful fields include owner, purpose, authority level, control tier, affected population, current version, supplier, production volume, exceptions, incidents, realised benefit, next review and expiry. The board should see changes in exposure and evidence quality, not a count of agents.

An agent's authority should expire unless renewed. Re-authorisation forces the owner to confirm that the use case, controls, evidence, economics and responsible people remain current. Dormant or ownerless agents should be disabled.

13. Use staged deployment as an evidence programme

A production programme should move through discovery, simulation, shadow operation, bounded pilot and controlled scale. Discovery maps the process and consequence. Simulation tests representative and adversarial cases without business-system action. Shadow operation compares the agent's proposed decisions with the existing process. A bounded pilot allows restricted action for a defined population and period. Scale follows evidence against pre-agreed gates.

The pilot population should represent the intended deployment. Easy cases can prove integration while overstating productivity and quality. The test set should include frequent cases, material exceptions, different languages or customer segments, poor-quality data, supplier outages and malicious inputs where relevant. Failure handling should be tested as deliberately as normal flow.

Entry and exit criteria should be approved before the pilot. These can include minimum evidence completeness, quality by material segment, exception rate, control performance, reviewer

capacity, incident readiness, user adoption and fully loaded unit cost. Averages should not conceal a severe outcome in a small group.

The investment gate should distinguish learning capital from scale capital. Early spending builds process knowledge, data, integration and control. Larger commitments follow evidence of repeatable outcomes and manageable residual risk. This reduces the pressure to justify sunk cost by broadening deployment prematurely.

The programme should retain adverse evidence. Failed cases, overrides, security tests and user complaints inform the boundary. A pilot is successful when it reveals whether and how the operating model can work, including a decision to narrow or stop the use case.

Table 4. Deployment and capital gates

Stage	Purpose	Required evidence	Decision
Discover	Define outcome, process and consequence	Baseline, mandate, owners, obligations and alternatives	Proceed to controlled testing or redesign
Simulate	Test capability and failure without live action	Representative cases, adversarial tests and control design	Enter shadow operation or restrict scope
Shadow	Compare proposed actions with current decisions	Segment quality, disagreement analysis and reviewer findings	Approve a bounded pilot or remediate
Pilot	Demonstrate live operation within limits	Outcomes, control performance, incident readiness, adoption and unit cost	Scale, hold, narrow or stop
Scale	Expand population or authority deliberately	Stable results, financial conversion, capacity and assurance	Release staged capital and re-authorise
Retire	Remove obsolete or unsafe capability	Data, access, records, continuity and stakeholder plan	Revoke identity and close obligations

Each gate requires current evidence for the tested configuration and population.

14. Apply the model to a hypothetical shared-services programme

Consider a hypothetical company processing 480,000 annual cases across accounts payable, procurement, customer operations and workforce administration. The author assumes annual relevant labour and external-service cost of USD 24 million. Average end-to-end cycle time is four business days, first-time-right performance is 91 percent and 12 percent of cases require supervisory review. These are analytical assumptions.

Management proposes an agentic workflow that reads incoming material, checks approved records, classifies the case, requests missing information, prepares an action and executes limited routine cases. It cannot create a supplier, change bank details, approve a payment, make an employment decision or issue a regulated customer communication. These actions require existing human authority.

The programme maps 62 percent of cases as standard, 25 percent as judgment cases and 13 percent as exceptions. In simulation, the agent completes 88 percent of standard cases correctly under the defined test, prepares usable recommendations for 70 percent of judgment cases and routes the remainder. The figures are assumed and do not establish real-world performance.

The decision-rights design places standard data retrieval and classification at Observe, evidence-backed recommendations at Recommend, draft records at Prepare and a small set of reversible administrative updates at Execute. Commit remains human-controlled. The agent has a unique identity, read access to specified systems, typed tool calls, counterparty allowlists, rate limits and no access to credentials in its prompt context.

The pilot begins with 20,000 cases. Management assumes that average human handling time falls from 18 minutes to 11 minutes, while oversight and exception effort add two minutes per case across the population. The net assumed release is five minutes per case, or roughly 1,667 hours for the pilot. Finance values only capacity that has an approved conversion plan. Half supports increased volume, one quarter avoids contractor spend and one quarter remains unconverted during the pilot.

Quality is measured by case type, customer effect and downstream correction. The pilot requires first-time-right performance of at least the existing 91 percent in every material segment, zero unauthorised commitments, complete logs, successful rollback tests and a defined response for severe incidents. A breach of access, an incorrect high-consequence action or loss of audit evidence stops the affected workflow.

At the end of the pilot, the board does not ask how many tasks the agent performed. It asks whether the operating outcome improved, whether authority stayed within the mandate, whether people could challenge and stop the system, whether adverse cases were understood and whether capacity converted into value. Scale is approved only for case types that meet these tests.

15. Build the investment case around option value and control cost

The financial model should separate fixed implementation cost, variable operating cost, control cost and expected benefit. Fixed cost includes process redesign, data work, integration, evaluation, security, training and change. Variable cost includes model inference, cloud, software, monitoring, human review and support. Control cost may rise with consequence and autonomy.

Benefits can include avoided hiring, lower external spend, increased volume, reduced error, faster cash conversion, improved service and reduced expected loss. Each benefit needs a baseline, measure, owner, attribution method and timing. Management should avoid counting released hours and headcount reduction for the same capacity.

Scenario analysis should cover adoption, quality, model price, review load, demand growth, supplier change and remediation. The downside should include a period of manual fallback, customer redress and delayed scale. The model should show which assumptions drive value and which can be tested during the next gate.

Option value matters. A reusable identity, evaluation, logging and tool-control layer can reduce the cost of future use cases. Process knowledge and clean data can create value even when a particular agent is stopped. These benefits should be described separately from realised cash because they are uncertain until used.

Capital release should follow evidence. The first tranche funds discovery and simulation. The second supports integration and a bounded pilot. Scale capital follows verified control and

economics. This makes the programme easier to stop, redirect or expand without treating the original business case as a permanent commitment.

16. Execute a ninety-day operating-model agenda

During the first 30 days, management should inventory current agents and AI-enabled workflows, identify accountable owners and select one outcome with measurable baseline data. The team should map the process, affected population, applicable obligations, existing controls and alternatives. It should draft the authority mandate and consequence classification.

During days 31 to 60, the team should configure separate identity, tools and limits; build representative evaluation cases; test security and failure scenarios; define the oversight role; and agree the productivity bridge with finance. The incident team should practise pausing, revoking and recovering the workflow. Employee and stakeholder engagement should follow the use case and applicable requirements.

During days 61 to 90, the programme should run shadow operation or a bounded pilot. Management should review quality by segment, tool use, exceptions, override patterns, review capacity, user adoption and fully loaded cost. It should record the financial conversion of any released capacity and retain adverse evidence.

The day-90 decision should be explicit: scale, hold, narrow, redesign or stop. Approval should name the next authority boundary, population, capital, owner, metrics and review date. The agent's identity and mandate should expire if approval is not renewed.

The broader enterprise programme can then reuse the governance and technical components. Each new workflow still requires its own outcome, consequence analysis, evidence and accountable owner. A common platform reduces friction; it does not make every use case equivalent.

17. Recognise limitations and open questions

Evidence on workplace generative AI is growing, while evidence on autonomous, tool-using agents in sustained production remains limited. Results from assistants, selected occupations or vendor pilots cannot be transferred mechanically to a different workflow. Organisational context, task mix, data, incentives, regulation and implementation quality influence outcomes.

Productivity measurement can be confounded by novelty, selection, changing demand and concurrent process improvement. Longer-term effects on learning, judgment, workforce structure and service quality may differ from a short pilot. Firms should continue measurement after scale and protect the ability to compare outcomes over time.

Human oversight has limits. Reviewers can become overloaded, defer to the system or lose the skills needed to identify error. Increasing apparent explainability can create confidence without improving correctness. Oversight design should therefore combine people with external limits, independent validation, sampling and stopping mechanisms.

Rules and standards continue to develop. Organisations operating across jurisdictions should obtain current legal advice for their use cases. They should record which obligations and guidance informed each deployment and update the analysis when the law, system or context changes.

The hypothetical example in this paper is a decision aid. Its values are assumptions, and its control design requires adaptation to the organisation's actual systems, risks, contracts and workforce.

18. Conclusion

Agentic AI turns operating-model design into a delegation problem. An organisation must decide which actions a system may observe, recommend, prepare, execute or commit. It must connect that authority to consequence, reversibility, evidence, access, oversight and recovery.

The strongest programmes begin with a measurable outcome and redesign the complete workflow. They give each agent a unique identity and enforce its mandate outside natural-language instructions. They build human oversight as a real operating capacity, preserve evidence, test failures and scale through explicit gates.

Productivity is demonstrated when improved work converts into a financial or service outcome while quality and control remain within appetite. A count of automated tasks cannot establish that result. Finance, operations, technology, risk and the affected workforce each hold part of the evidence.

Boards can use the framework as a practical test. Who owns the outcome? What may the agent do? What can it affect? Which evidence supports the action? How is error detected and contained? Can a person challenge and stop it? How does released capacity create value? Autonomy should expand only when the answers remain current and verifiable.

References

1. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

2. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

3. National Institute of Standards and Technology AI Resource Center. AI RMF Core. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

4. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, including Articles 14 and 26. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

5. European Union. Regulation (EU) 2016/679, Article 22. https://eur-lex.europa.eu/eli/reg/2016/679/oj

6. UK Department for Science, Innovation and Technology. AI Management Essentials tool. Updated 2026. https://www.gov.uk/government/consultations/ai-management-essentials-tool/ai-management-essentials-tool-accessible

7. UK National Cyber Security Centre. Thinking carefully before adopting agentic AI. 2026. https://www.ncsc.gov.uk/blogs/thinking-carefully-before-adopting-agentic-ai

8. UK National Cyber Security Centre and international partners. Guidelines for secure AI system development. 2023. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

9. National Institute of Standards and Technology. Strengthening AI Agent Hijacking Evaluations. 2025. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

10. UK Government Communication Service. The People Factor: A human-centred approach to scaling AI tools. 2025. https://www.gov.uk/government/publications/a-human-centred-approach-to-scaling-and-de-risking-ai-tools/the-people-factor-a-human-centred-approach-to-scaling-ai-tools-html

11. OECD, BCG and INSEAD. The Adoption of Artificial Intelligence in Firms: New Evidence for Policymaking. 2025. https://doi.org/10.1787/f9ef33c3-en

12. Brynjolfsson, E., Li, D. and Raymond, L. R. Generative AI at Work. Quarterly Journal of Economics, 140(2), 2025, 889-942. Working-paper record: https://www.nber.org/papers/w31161

13. Gmyrek, P. et al. Generative AI and Jobs: A Refined Global Index of Occupational Exposure. ILO Working Paper 140. 2025. https://doi.org/10.54394/HETP0387

14. Milanez, A., Lemmens, A. and Ruggiu, C. Algorithmic Management in the Workplace: New Evidence from an OECD Employer Survey. OECD AI Papers No. 31. 2025. https://doi.org/10.1787/287c13c4-en

15. International Organization for Standardization. ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. https://www.iso.org/standard/81230.html

16. Information Commissioner's Office. Guidance on AI and data protection. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/

17. Information Commissioner's Office. AI and data protection risk toolkit. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/ai-and-data-protection-risk-toolkit/

18. National Institute of Standards and Technology. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2 E2025. https://csrc.nist.gov/pubs/ai/100/2/e2025/final

19. National Institute of Standards and Technology. Lessons Learned from the Consortium: Tool Use in Agent Systems. 2025. https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems

20. Singapore Infocomm Media Development Authority and AI Verify Foundation. Model AI Governance Framework for Generative AI. 2024. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/model-ai-governance-framework-for-genai

21. OECD. OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market. 2023. https://doi.org/10.1787/08785bba-en

22. OECD. How widespread is algorithmic management in workplaces? 2025. https://www.oecd.org/en/publications/how-widespread-is-algorithmic-management-in-workplaces_cda7a114-en.html

23. UK AI Security Institute. Will it become harder to oversee AI systems? https://www.aisi.gov.uk/blog/will-it-become-harder-to-oversee-ai-systems

24. US Equal Employment Opportunity Commission. Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures. 2023. https://www.eeoc.gov/select-issues-assessing-adverse-impact-software-algorithms-and-artificial-intelligence-used

25. Committee of Sponsoring Organizations of the Treadway Commission. Achieving Effective Internal Control Over Generative AI. 2026. https://www.coso.org/generative-ai

About the Author

Authored by Chennakeshav Adya

Independent Researcher

Chennakeshav (CK) is a corporate finance and investment banking executive with 25+ years of global experience in deal origination, structuring and execution across M&A, growth capital and corporate strategy. He has led value-creation mandates for founders, corporates and funds, bridging the boardroom view to hands-on execution and close.

His career spans Morgan Stanley, HSBC, Lloyds Banking Group, EWEC, ADQ portfolio companies and Emirates Growth Fund, across TMT, real estate, fintech, deeptech, cleantech, infrastructure and energy. He has partnered with C-suite leaders, private equity and venture funds, sovereign wealth funds and family offices to finance complex fund raises and scale-up ventures, and has led M&A due diligence, post-merger integration and business-transformation initiatives to create value.

At Matchpoint Partners he is Managing Partner, leading the firm's corporate finance, M&A and capital-raising practice. He holds an MBA from London Business School, an engineering degree from VTU and a Master of Laws (LLM, in progress) from UCL London.

An active start-up mentor, CK mentors at Techstars, DIFC FinTech Hive, Startup Grind, Founder Institute and IN5, serves as Entrepreneur Mentor in Residence (EMiR) at London Business School, and judges the Entrepreneurship World Cup.