

AI Governance Roundtable #6: Governing Agentic AI

How can and should organisations govern agentic AI? What needs to be understood by whom, when, and why?

*This is the **sixth** in a series of roundtables convened by the NUS AI Institute's AI Governance and Policy team—building on work done under the auspices of AI Singapore—for representatives from industry, government, and academia to discuss responsible AI. Such discussions are typically too narrow and too broad. Too narrow in that a few voices dominate the discussion—notably those in the United States and Europe, with China sometimes included. Too broad in that discussion is often limited to generalities and principles. This project aims to address both aspects of this problem, involving a wider set of stakeholders—in particular those from Southeast Asia—in more focused discussions of specific challenges in the application of responsible AI to particular questions.*

Rapporteur: Dr. Eric J. W. Orlowski

Location: IBM, Singapore

Date: 12 June 2026

Executive Summary

How can organisations move from experimenting with agentic AI to deploying it responsibly? This was the topic of a roundtable organised by the NUS AI Institute, and hosted by IBM Singapore, synthesised in this report. Participants from industry, government, academia, civil society, law and finance examined systems that do more than generate outputs: they can plan, invoke tools, access data, trigger workflows and act with varying autonomy on behalf of people and organisations.

The discussion repeatedly returned to a foundational shift. Agentic AI moves the governance task from reviewing outputs towards controlling behaviour at runtime. Principles remain necessary, but cannot determine which data or tools an agent may access, which actions it may take, when a human must intervene or who bears responsibility when harm occurs. Responsible deployment therefore requires principles to be translated into operational governance, technical controls and clearly assigned human and organisational accountability. These arrangements should be proportionate to the use case, autonomy, reversibility and consequences of failure.

The roundtable identified four governance gaps across three connected domains. Technical and operational gaps concern agent identity, permissions, traceability and forms of human oversight capable of working at speed and scale. Legal and institutional gaps arise from uncertainty about how existing laws and duties apply across complex value chains, how liability should be calibrated, and which actors and authorities must respond when incidents cross organisational or regulatory boundaries. Organisational and workforce gaps concern whether enterprises possess the governance capacity, skills, confidence and employee participation needed to adopt agentic systems without eroding human judgement or institutional resilience. These enterprise-level challenges sit

alongside systemic risks—including cybersecurity, disinformation, financial disruption and critical-infrastructure failures—that may propagate across sectors and borders.

The report proposes five recommendations for action. Organisations and regulators should:

1. Establish proportionate agent identity, permissioning and accountability infrastructure;
2. Replace generic “human-on-the-loop” requirements with use-case-specific oversight models;
3. Clarify the application of existing law while developing calibrated liability and coordinated incident-response mechanisms;
4. Treat deployment as an organisational transformation programme supported by sandboxes, assurance testing, workforce participation and AI literacy;
5. Strengthen cross-sector and regional resilience for spillover risks.

Singapore’s pragmatic, targeted approach provides a useful foundation, but voluntary frameworks and existing law require clearer operational and regulatory articulation if organisations are to deploy with confidence. For agentic AI, this should include sector playbooks, assurance sandboxes, agent identity pilots, incident-response protocols and procurement expectations that define the minimum governance capabilities providers and platforms must expose by default. Responsibility must be shared without becoming diffuse: regulators should clarify expectations, enterprises should govern deployments before they scale, providers and platforms should supply the technical and operational controls on which deployers depend, and academia, civil society and regional bodies should contribute research, scrutiny and coordination. Governance, approached in this way, becomes not merely a constraint on adoption but part of the infrastructure for responsible adoption, preserving accountability, trust, human capability and public resilience.

About the Roundtable: Governing Agentic AI

This roundtable, organised by the NUS AI Institute and hosted by IBM Singapore, focused on the responsible deployment of agentic AI in enterprise settings. Agentic systems are able not only to generate outputs in response to prompts, but also to plan, take multi-step actions, invoke tools, interact with other systems and act with varying degrees of autonomy on behalf of users or organisations. This creates opportunities for productivity, service delivery, workflow automation, customer engagement and organisational transformation. It also raises questions of accountability, oversight, liability, technical control, workforce impact and trust, particularly when systems connect to enterprise infrastructure or act across organisational workflows. The session brought together participants from industry, government, academia, civil society, law and finance.

The discussion was framed around a central question: how can organisations in Singapore move from experimentation with agentic AI towards responsible deployment? Participants used three interconnected layers of governance as a starting point for the discussion: principles and ethics provide a strategic “north star”; operational governance assigns responsibility and translates policy into everyday practice; and technical governance supplies the safeguards, monitoring, access controls, testing procedures and runtime controls required once systems are deployed. A recurring premise of the discussion was that principles remain essential, but acquire practical force only when translated into organisational and technical arrangements. As one participant put it: “Rules are fine, but how do you know that the rules are actually being implemented by your AI systems? Especially your Agentic AI systems; especially when they’re spread all across your organisation?”

Singapore has not adopted a horizontal AI law, relying instead on targeted interventions in particular areas of harm, practical guidance, industry engagement and co-developed governance frameworks.

For agentic AI, this posture should be translated into the operational mechanisms set out below, with responsibilities and controls made explicit. Governance can then function as enabling infrastructure for confident adoption rather than merely as a brake on it. “Agentic AI” should not, however, be treated as a uniform category. Risk varies with the use case, degree of autonomy, access to data and tools, capacity to transact, reversibility and potential impact. A low-risk internal productivity agent does not warrant the same controls as a system that initiates financial transactions, influences hiring, manages critical infrastructure or interacts with public-facing platforms.

Following the Chatham House Rule, this report synthesises the issues participants regarded as most urgent. It does not attempt to provide a comprehensive legal or technical framework. It first explains the shift from governing systems that produce outputs to governing systems that act, then examines the resulting governance gaps, systemic risks, and recommendations for the different stakeholders.

Foundational Shift: From Output Review to Runtime Control

Earlier waves of AI governance were largely concerned with systems that classified, ranked, predicted, recommended or generated content. These systems could cause significant harm, particularly when they shaped access to employment, credit, healthcare, information, public benefits or other services. The governance task nonetheless tended to centre on whether an output was sufficiently reliable, fair, explainable and appropriate for its intended use. Agentic AI changes the problem because an agentic system may not only produce an answer but take steps to achieve a goal. It may call external tools, access enterprise databases, trigger workflows, communicate with other systems or carry out tasks on a user’s behalf. Enterprise applications may include updating records, conducting procurement, scheduling operations, initiating customer-service actions or supporting financial and legal workflows. More sensitive deployments may affect employment, regulated transactions, public services, critical infrastructure or security.

An erroneous output and an erroneous action create different governance problems. A flawed output may mislead someone or require correction. A flawed action may change a system state, release information, move money, deny access or trigger downstream processes. Organisations must therefore ask not only whether an output was correct, but whether an action was authorised, whether the system was permitted to take it, whether it can be reversed and who remains accountable if harm occurs. Established concerns such as bias, hallucination, opacity, privacy and security remain relevant, but now arise within systems capable of acting across workflows and digital infrastructure. High-level principles cannot specify which tools an agent may access, which data it may use, which actions it may take or when humans must intervene. Those decisions require procedures, permissions, technical controls, assigned roles, documentation, escalation pathways and incident-response mechanisms. This growing impetus to govern these emergent agentic AI risks was summarised by one participant as “and then it will run, and it will fill in the gaps, and continue to run, and the more we start doing these things, the more opaque our understanding of what’s happening becomes”.

Agentic AI therefore shifts governance from output review towards runtime control: defining and enforcing the bounds within which a system may act, retaining evidence of its actions and preserving the ability to intervene or recover. This is not a purely technical undertaking. Legal, compliance, cybersecurity, risk and domain expertise are needed to decide what those bounds should be and who may change them. Governance must also begin earlier. When agentic capability arrives through software-as-a-service (SaaS), cloud platforms or software updates, deployers are limited to the controls providers expose. Decisions about acceptable use and human authority must therefore be made during design and procurement, tested before deployment and revisited as capabilities or operating conditions change. Incident response and decommissioning complete the same lifecycle.

The intensity of governance should reflect the use case. Lower-risk systems may operate internally on reversible tasks subject to meaningful review. Systems involving sensitive data, external users, regulated transactions, safety-critical functions or consequential decisions need stronger controls. Assessment must account for autonomy, reversibility and the consequences of failure rather than treating “agentic AI” itself as a risk tier. Singapore’s targeted governance posture offers a useful starting point: existing laws, sectoral rules, governance frameworks, assurance mechanisms and regulatory practices may be adapted before entirely new law is considered. Reliance on existing frameworks, however, does not remove the need for greater clarity about responsibility, due diligence and governance expectations. Agentic AI changes where governance must operate—within the design, deployment, operation and monitoring of systems that can act.

Governance Gaps

The roundtable identified four governance gaps across three connected domains: technical and operational governance; legal and institutional governance; and organisational and workforce governance. Although analytically distinct, these domains interact in practice. Controls depend on clear authority, legal responsibility depends on operational evidence, and neither can function without organisational capacity.

Gap 1: Agent Identity, Permissioning and Accountable Authority

Agentic AI introduces an identity and accountability problem. In conventional digital systems, organisations generally know which user, account, system or service is accessing a resource. In regulated contexts, this is not merely a technical convenience: organisations must know which systems are authorised to access data, invoke tools, trigger workflows or conduct transactions, and which people or entities remain responsible. Agents complicate this picture because they may act for users, organisations or other systems while their origin, purpose, authority and permissions remain unclear. The issue is particularly acute in financial services. Institutions may be technically able to support transactions, account opening, foreign exchange and anti-money-laundering workflows, yet unable to serve an agent without knowing whom it represents, whether it is authorised and who can be held accountable. “Know Your Agent” may therefore extend “Know Your Customer”, as “not only do you need to have agent identity, you also need to have human accountability. And we need to connect the two”. However, this cannot simply reproduce KYC processes. Platforms face related questions when distinguishing human users, authorised agents, commercial bots and malicious or unidentified automated systems.

This is not only a question of labelling. Trust requires evidence of an agent’s provenance, approved purpose, scope of authority and relationship to an accountable principal. Those attributes may change when an agent is updated, reassigned or permitted to call sub-agents, so identity must be maintained across the deployment lifecycle rather than established once at registration. Agent registries could record which agents exist, what they are intended to do, which permissions they hold and who owns them. Yet a registry is not a complete solution. Large numbers of persistent, temporary or task-specific agents and sub-agents could make individual registration unworkable, producing records too dynamic to support meaningful governance while creating an illusion of control. Registries may also need to operate across organisational boundaries. An enterprise can inventory agents it deploys directly, but may know less about agents embedded in vendor products, acting through platforms or interacting with its services from outside. High-risk contexts may therefore require interoperable credentials or attestations that communicate authority, risk-relevant attributes

and accountable ownership without disclosing unnecessary information. These arrangements must also support revocation and change: a trustworthy identity system must show when authority has expired, permissions have been reduced or a previously approved agent has materially changed.

A proportionate identity, permissioning and accountability layer is therefore likely to be more practical. High-risk agents may require individual registration, while lower-risk or short-lived agents may be governed through registered system classes, use cases, owners, deployment environments or permission structures. Existing mechanisms—authentication, authorisation, interface contracts, access controls and non-repudiation—provide a foundation, adapted for systems that are more autonomous and context-sensitive than conventional software services. The aim is sufficient visibility and control to support operational trust, rather than bureaucratic completeness. Identity must remain linked to human or organisational accountability. An agent is not a legal person or a place for responsibility to disappear. Where agents invoke other agents, third-party models, external tools or platforms, responsibility may become fragmented across developers, deployers, operators, users and providers. That complexity makes traceability more important. Enterprises therefore need to know which agents are operating, what authority they hold and who remains accountable for them. Providers and platforms must make this information machine-readable and enable customers to distinguish authorised agents from unidentified or malicious automation. Regulators may also need to specify the evidence required in high-risk settings.

Gap 2: Human Oversight for Autonomy, Speed and Scale

Human oversight is among the most frequently invoked principles in AI governance, but also among the least precise. The presence of a person somewhere in a workflow is insufficient if that person rubber-stamps recommendations, receives more alerts than can be reviewed, lacks the information needed to judge the system or bears responsibility without authority to intervene. One participant framed it as “what is it this agentic AI system is doing for them? Can they understand it qualitatively? If they disagree, can they override it? Can they disable it?” In other words, oversight must be designed around points at which human judgement can genuinely shape outcomes.

Continuous and time-sensitive systems make pre-action approval especially difficult. Participants discussed utilities and grid balancing, where parameters change rapidly, as well as manufacturing environments in which agents may coordinate workflows across machines and production stages. Requiring approval before every action may be operationally impossible or defeat the purpose of the system. Meaningful oversight may instead require experts to define operating boundaries, approve the scope of autonomy, set escalation thresholds and establish fallback and recovery mechanisms. Professional and safety-relevant work poses a related problem. In engineering, legal, medical, financial and other specialised contexts, fluent and rapid performance does not establish competence. Organisations must determine which decisions still require professional judgement and provide users with enough information to assess reliability, limitations and appropriate use. The relevant explanation may be functional rather than technical: users need to understand what the system is doing, how it affects them and whether an action can be contested, overridden or disabled.

Oversight may take the form of approval before action, monitoring while a system operates or post hoc review for low-risk, reversible actions. Time-sensitive agents may operate within tested limits without immediate approval if humans retain practical powers to pause, override, redirect and escalate. Requirements to review every action or log may overwhelm supervisors and merely simulate control. Using one agent to audit another may assist at scale, but adds another system whose reliability and permissions must be governed. Meaningful oversight consequently involves more than selecting a point in a workflow at which a person clicks “approve”. Organisations must decide which information reaches an overseer, how alerts are prioritised, how much time is available to respond

and whether intervention can actually prevent or reverse harm. Domain experts are essential in setting these thresholds because technical teams alone may not recognise when an apparently routine action becomes professionally, legally or operationally consequential. Accountability should follow practical authority: individuals should not be designated as responsible for systems they cannot understand, stop or change.

Effective oversight also depends on preserving human capability. As tasks are delegated, employees may have fewer opportunities to practise the skills needed to understand, challenge or recover from agentic decisions. This is both a workforce concern and an organisational resilience concern, developed further below. The immediate test is not whether a human is nominally present, but whether human judgement remains effective where it can prevent harm, support trust and preserve accountability. Oversight must also be workable for people affected through systems they do not control. Smaller businesses and other platform users may not need explanations of model architecture, but they require notice of consequential agentic activity and practical routes to request review or redress. Otherwise, formal human oversight inside the deploying organisation may coexist with little meaningful agency for those outside it.

Gap 3: Liability, Institutional Responsibility and Regulatory Certainty

Harms from agentic AI may involve model developers, agent developers, providers, deployers, operators, users, platforms, data providers, APIs and third-party tools. When systems act across workflows, invoke other systems or delegate tasks to other agents, it becomes difficult to determine where responsibility begins and ends. The problem is sharpened because agents may initiate transactions, change system states or make decisions affecting individuals, businesses and public infrastructure rather than merely provide advice. Technology-agnostic laws—including data protection, cybersecurity, consumer protection, employment law, sectoral regulation, contract law and civil liability—may already apply to many harms. The central gap is often uncertainty about how obligations attach across a distributed value chain: which actors bear particular duties, which regulator has jurisdiction and what constitutes reasonable due diligence when control is shared among developers, providers, deployers, platforms and users.

This uncertainty may deter adoption, particularly among micro-, small- and medium enterprises (MSMEs) and less AI-mature organisations without the legal and technical resources to interpret fragmented obligations. Organisations are unlikely to deploy confidently if they cannot assess their exposure when something goes wrong. Indeterminate liability may also burden actors unable to control downstream use. Greater legal and regulatory certainty is therefore needed both to protect against harm and to help organisations distinguish manageable risks from deployments they should not undertake.

A binary choice between fault-based and strict liability is unlikely to suffice. Fault-based approaches preserve room for innovation but may be difficult for affected parties to use when causation and responsibility are opaque. Strict liability may offer stronger protection for serious harms or where affected parties have little bargaining power, but may deter useful deployment if imposed too broadly or on actors unable to control downstream use. Liability should instead reflect the seriousness of potential harm, degree of autonomy, reversibility and control exercised by each actor. Roundtable participants cautioned, however, that such calibration is easy to state and difficult to implement without clearer law and regulation: “the risk is we’re still quite far away from having a comprehensive legal framework that addresses these technologies”, one noted; while another noted that “I’m, in fact, in favour of more strict laws and certainty, so that these people know if you are this size, then these are the laws that you follow. There’s certainty in that; the human who’s accountable can say ‘well, I’ve done my best. I followed the law’”.

Recognised governance practices can help make calibration operational. Risk assessment, testing, permission controls, oversight, monitoring, documentation and incident planning can provide evidence that an organisation exercised due care. Compliance with a framework should not automatically absolve an organisation where harm was foreseeable or controls were inadequate. It may nevertheless help distinguish meaningful precautions from box-ticking or deployment without adequate understanding. Clarity is required at both the general and sectoral levels. Cross-cutting guidance can explain how existing duties attach to developers, providers and deployers, while sector regulators specify what due diligence and redress require in contexts such as finance, employment, healthcare, public services and critical infrastructure. Coordination matters where a single failure engages several regimes at once. Without an agreed route for notification and lead-agency responsibility, organisations may face duplicative requirements while affected parties still struggle to obtain a timely response.

Legal responsibility must also connect to contracts and operational response. Enterprise systems commonly combine proprietary models, open-source components, cloud infrastructure, APIs, platforms, third-party tools and internal workflows. Contracts should allocate responsibility for performance, integration, data handling, security, monitoring, warranties, documentation and post-incident support. Contracts should ensure that deployers can obtain the evidence and operational support needed to investigate, contain and reverse failures. Incident arrangements must specify who can pause the system, which authorities must be notified, what evidence must be preserved, how downstream actions can be contained or reversed and who communicates with affected parties. The gap is therefore not only legal uncertainty, but an insufficient connection between legal responsibility, governance practice, contractual allocation and operational response.

Gap 4: Organisational Readiness, Workforce Trust and Ecosystem Inclusion

Technical capability alone does not provide the governance processes, workforce knowledge, institutional confidence or operational controls required for responsible deployment. Organisations may be uncertain which models to use, how to manage data-sharing risks, how to assess cost and vendor dependence or how to govern systems once deployed. High-profile accounts of unpredictable agent behaviour may also discourage experimentation. Across Southeast Asia, differences in digital maturity, technical capacity and bargaining power make these barriers especially significant for MSMEs. Large enterprises may be able to build internal governance teams, negotiate contractual protections and conduct independent testing. Smaller organisations often have fewer technical staff, weaker leverage with vendors and platforms and less capacity to absorb a failed deployment. They may also receive agentic functions as updates to products already in use rather than through an explicit procurement decision. In such cases, the provider or platform controls the identity, permissioning, logging, data-use information and intervention functions available to the customer. Responsible adoption therefore depends partly on an ecosystem of shared resources and on providers making safe configuration and meaningful choice available by default, rather than assuming every customer can construct governance from scratch.

Sandboxes and assurance testing can turn abstract uncertainty into informed judgement. Participants described structured experiments that allowed initially hesitant organisations to observe system behaviour, identify concrete risks, evaluate mitigations and build confidence before wider deployment. A key takeaway from previous experience, it was noted, was “that the government realised that they had to get their hands dirty and experiment with the technology and see for themselves. What are the issues with agents, and how can they mitigate them before they decide whether they want to adopt?” The participant continued: “It doesn’t mean we can mitigate all the risks, but there are a lot of things they can do to really minimise the risks”. Their purpose is not to

prove that risk has been eliminated, but to expose context-specific trade-offs around access, permissions, integration, oversight and security that principles or generic guidance cannot reveal.

Readiness also depends on AI literacy and workforce trust across leadership, legal, compliance, HR, operations, procurement, customer service, product teams and frontline functions. If understanding remains concentrated among specialists, organisations will struggle to embed governance in everyday practice. Concerns about displacement, deskilling, surveillance and loss of professional judgement must be addressed. Examining change at the level of tasks, rather than whole jobs, helps distinguish activities that may be automated or augmented from those that depend on human judgement, creativity, empathy or professional responsibility. Employees and domain experts should help identify use cases, define success, test behaviour and determine appropriate oversight. Their participation is both procedurally important and practically useful: they know where workflows fail, where contextual judgement matters and where automation may create operational or customer-facing risks. Excluding these perspectives can produce technically capable systems that are poorly aligned with actual work and mistrusted by those expected to use them. Organisations therefore need channels through which employees can report failures, contest actions and challenge unsuitable deployments.

Many MSMEs will encounter agents through platforms they neither procure nor control. Agents embedded in sales, logistics, payments, advertising, hiring or customer-service platforms may affect pricing, rankings, discounts, visibility, customer allocation or business strategy. Affected businesses may not need details of model architecture, but they do need functional transparency about what the system does, how it affects them, whether its actions can be contested or overridden and how risks and benefits are distributed. Without such controls, agentic AI may deepen platform dependence rather than strengthen smaller businesses. Responsible deployment is therefore an organisational transformation process involving governance capacity, experimentation, procurement, workforce participation, platform accountability and ecosystem support. It must preserve the human skills and institutional knowledge needed to evaluate, challenge and recover from agentic decisions. An organisation may otherwise become more efficient while becoming less able to understand or control its own operations, creating new forms of institutional fragility.

Cross-Cutting Risk: Public and Systemic Spillovers

Enterprise deployments may create risks beyond the deploying organisation, particularly in cybersecurity, disinformation, finance, critical infrastructure, law enforcement and regional security. Agentic systems increasingly connect organisations to wider digital, economic and social infrastructures, allowing decisions and failures to propagate across sectors and jurisdictions. Agentic AI can automate elements of cyber operations, including reconnaissance, vulnerability discovery, social engineering, code generation and attack adaptation. Legitimate but poorly configured agents—with excessive permissions, weak authentication, unsafe tool access or inadequate monitoring—may also expand an organisation’s attack surface. Shadow agents deployed without approval, documentation or oversight create risks that security and governance teams may not know exist. Cybersecurity policy must therefore address agent configurations alongside users, devices, applications and conventional software services.

Agentic systems may also lower the cost of generating, targeting, testing and distributing disinformation or harmful content. They may adapt campaigns dynamically, interact with users and coordinate activity across languages, platforms and jurisdictions. Existing detection and takedown mechanisms may be strained, requiring coordination among platforms, regulators, cybersecurity agencies, civil society and regional institutions rather than reliance on domestic content-governance measures alone. The regional dimension is material. Southeast Asia’s linguistic diversity and uneven

institutional capacity may make harmful campaigns difficult to detect consistently, while content, infrastructure and financial flows cross borders quickly. Information sharing must therefore include context about local languages and political conditions, not merely technical indicators. Civil society and independent researchers can help identify harms that are visible to affected communities before they become legible to platforms or authorities.

Financial and critical-infrastructure applications raise concerns about speed, scale and cascading effects. Misconfigured or malicious agents capable of initiating payments, executing microtransactions or interacting with financial systems could cause disruption faster than institutions can intervene, echoing the dynamics of flash crashes. As one participant asked: “One thing is the scale at which this happens. If it explodes too quickly, will we lose our ability to control it?” Failures in utilities, transport, communications and energy may affect the public, other sectors and national resilience even where the deploying organisation can manage its immediate system. Governance must therefore consider how failures may propagate beyond the enterprise. Law enforcement, cyber operations, autonomous systems and military use similarly risk accelerating action, compressing deliberation and obscuring attribution. Malicious actors will not observe responsible-use guidelines; preparedness therefore requires investigative capacity, operational safeguards, international cooperation and regional dialogue. These concerns may seem distant from enterprise adoption, yet the underlying capabilities and infrastructure frequently overlap.

Auditability cuts across all these risks. High-volume logs and intermediate decisions may be too complex to support investigation, while agents auditing agents introduce further reliability and governance questions. The objective is usable auditability: records detailed enough for investigation and accountability but structured so that human experts and institutions can interpret them. Strong enterprise governance is necessary, but systemic resilience also requires cross-sector and regional coordination. These spillovers complicate the boundary between private governance and public preparedness. Enterprises remain responsible for securing and monitoring their deployments, but no single organisation can evaluate every systemic dependency or coordinate a multi-sector response. Regulators and public agencies need mechanisms for aggregating incident information, identifying common failure patterns and running scenarios that reveal where individually reasonable deployments may create collective vulnerability. Such coordination should not wait for a major incident to demonstrate that fragmented oversight was inadequate.

Recommendations: Operationalising Singapore’s Targeted Approach

Drawing on the roundtable, the report proposes five mutually reinforcing recommendations for putting Singapore’s targeted governance approach into practice: sector playbooks, assurance sandboxes, agent identity pilots, common incident-response protocols and procurement expectations. Each should be calibrated to the sector, use case, degree of autonomy, reversibility and potential harm.

Many organisations will encounter agentic AI through SaaS products, cloud platforms and software updates rather than by building or separately procuring standalone agents. Their ability to govern these systems therefore depends partly on what providers and platforms enable them to see and control. At minimum, customers should be able to identify an agent and the party accountable for it; limit or revoke its authority; understand the data, models and tools on which it relies; reconstruct its actions and tool use from usable records; pause, override or roll back its activity; and obtain timely support when failures occur. This baseline should apply to embedded functionality as well as standalone agents and should not be confined to premium enterprise tiers. Platforms should also

distinguish agent-mediated from human activity and provide proportionate routes for suspension, contestation and redress.

Recommendation 1: Establish Proportionate Agent Identity, Permissioning and Accountability Infrastructure

Organisations and regulators should establish mechanisms through which agentic systems can be identified, authorised, monitored and connected to accountable parties. The objective is not to register every agent in the same way. Lower-risk internal systems may be governed through enterprise inventories recording owner, purpose, risk tier, model dependencies, permissions and escalation contacts. Higher-risk systems involving regulated transactions, sensitive data, consequential decisions, public-facing platforms or critical infrastructure may require more formal identity and authorisation.

Access should be limited to what the approved use case requires and reviewed as systems or contexts change. Agent permissions should, where possible, integrate with existing identity and access-management systems rather than rely on ad hoc arrangements. High-risk sectors should explore “Know Your Agent” approaches specifying what must be known before an agent can transact, access regulated infrastructure or act for a person or institution. Relevant information may include its owner, purpose, authorisation scope, audit trail, escalation contact and chain of legal responsibility. Providers and platforms should make an agent’s identity, provenance, current authority and accountable owner verifiable across organisational boundaries. Policymakers should avoid blanket registries that cannot operate at scale, while stating clear minimum expectations for high-risk settings. Depending on context, organisations may register individual agents, agent classes, systems, APIs, deployment environments or accountable owners.

Implementation should begin with discovery. Enterprises cannot govern agents they do not know are in use, including agentic functions embedded in existing software or introduced by employees outside approved procurement channels. Inventories should therefore connect technical discovery with an accountable owner, approved purpose, risk classification, dependencies and review date. Higher-risk agents should be subject to stronger authentication and non-repudiation, and their credentials should be revocable when ownership, purpose or system behaviour changes. Singapore could begin with agent identity pilots in high-trust settings such as financial services or enterprise procurement. These should establish what an organisation must verify before accepting instructions or transactions from an external agent.

Recommendation 2: Replace Generic “Human-on-the-Loop” Requirements with Use-Case-Specific Oversight Models

Organisations should choose oversight models according to risk, speed, reversibility and potential harm. Human-on-the-loop arrangements may suit continuous systems where people monitor behaviour and can pause, override, redirect or escalate. Bounded autonomy may suit systems that must act without immediate approval, provided they operate within tested limits supported by escalation thresholds, fallback mechanisms, rollback and incident-response plans. Low-risk and reversible actions may be suitable for post hoc review. Oversight should be designed before deployment and incorporated into use-case approval, procurement, system design and risk assessment. Regulators should specify what meaningful involvement requires rather than invoke a generic human-on-the-loop principle. Sector playbooks should define which oversight models are presumptively appropriate for common use cases and what evidence is needed to justify greater autonomy. Enterprises should determine what an agent can do, what could go wrong, whether its actions are reversible, who can intervene and what information overseers require. The answer should reflect the sensitivity of the data or decision and the feasibility of real-time intervention.

Providers and platforms should ensure that their intervention and record-keeping controls can support the chosen oversight model. Domain experts and employees should identify where professional judgement must remain decisive, while academic researchers evaluate how different arrangements affect intervention, automation bias and skill retention. Singapore's assurance sandboxes could test whether those arrangements work in realistic workflows: whether information reaches supervisors in time, can be understood and allows them to stop or reverse consequential actions. The results should inform system boundaries, staffing, training and decisions about which actions may remain automated.

Recommendation 3: Develop Calibrated, Risk-Based Liability and Incident-Response Mechanisms

Governments, regulators and legal researchers should map how existing technology-agnostic laws apply to agentic AI before determining where clarification, coordination, reform or new legislation is needed. The mapping should cover data protection, cybersecurity, consumer protection, employment, finance, safety, public-sector obligations, contract and civil liability. Sector regulators should then define expectations for approving high-risk uses, assessing risk, monitoring performance, responding to incidents and providing redress. Liability should reflect potential harm, autonomy, reversibility and each actor's control. Strict liability may be appropriate for serious risks to life, bodily integrity, critical infrastructure or public safety, while fault-based approaches may be more proportionate for lower-risk or reversible uses. Intermediate mechanisms—including presumptions of responsibility, insurance, heightened documentation and consideration of recognised governance practices—may also be needed.

Deployers and providers should document the information needed to allocate responsibility, investigate failures and support redress. Contracts should state which duties each party retains and what operational support will be available after an incident. Regulators should establish a common agentic-AI incident protocol, with sector-specific annexes, defining severity thresholds, notification routes, lead-agency responsibility and procedures for containment, evidence preservation and redress. Insurers and risk professionals can develop coverage models that reward meaningful testing and preparedness, while comparative research can identify legal uncertainty and workable allocations of responsibility across value chains. This mapping exercise should produce usable outputs rather than a catalogue of potentially relevant law. Enterprises need guidance on which party must perform particular checks, which records demonstrate due diligence, which contractual terms cannot displace statutory duties and where affected people can seek redress. Regulators should identify gaps created by agent-to-agent delegation, open-source components and cross-border services, and clarify which authority coordinates incidents spanning several regimes. Sandboxes and incident simulations can test whether proposed allocations work before they are relied upon in high-impact deployments.

Recommendation 4: Treat Responsible Deployment as an Organisational Transformation Programme, Not a Tool Rollout

Organisations should treat agentic AI deployment as a programme involving governance, workforce capabilities, procurement, risk management and culture. Sandboxes and assurance testing should precede deployment at scale where uncertainty is significant, allowing teams to observe behaviour, test safeguards, clarify oversight and determine whether a use case is appropriate. Participation should produce a documented decision on whether to deploy, the conditions attached to deployment and any residual risks requiring monitoring. AI literacy should extend across leadership, legal, compliance, HR, operations, procurement, product, customer-facing teams and frontline employees. Training should address not only how to use a system, but when to use it, what limitations require attention and how to escalate concerns. Task-level mapping should identify what can be

automated or augmented and what should remain human-led, providing a grounded basis for workforce planning and decisions about professional judgement.

Employees and domain experts should help identify use cases, define success, test systems and determine oversight. Organisations should provide channels for feedback, reporting, contestation, override and review so that failures can be surfaced before they become embedded. These mechanisms support procedural fairness, operational learning and workforce trust. MSMEs need shared sandboxes, templates, practical training, communities of practice and use-case repositories provided by regulators, industry associations, financial institutions, platforms and vendors. Platforms should provide usable controls and functional explanations. Academic researchers can evaluate organisational-readiness initiatives and study effects on work, skills, employee participation, institutional knowledge and organisational resilience.

Boards and senior leaders should set risk appetite and identify which uses require higher-level approval, but implementation must be cross-functional. Governance teams need authority to coordinate legal, security, procurement, HR, operations and domain functions rather than acting as an advisory layer after decisions have been made. Within Singapore, the common provider baseline could be translated into procurement expectations for public bodies and for regulated or high-impact deployments. Tender and contract documentation should specify which elements of that baseline the intended use requires. Workforce planning should track whether delegation changes the skills needed for oversight and recovery, not only whether it reduces time or headcount. Lessons from sandboxes and incidents should be retained and shared so that organisational knowledge grows as deployments scale.

Recommendation 5: Build Cross-Sector and Regional Resilience for Agentic AI Spillover Risks

Governments, regulators and organisations should develop resilience mechanisms for risks that individual enterprises cannot manage alone. Cybersecurity agencies and enterprise security teams should establish safe-configuration standards for authentication, authorisation, tools, data access, logging, monitoring, decommissioning and shadow agents, applying least privilege throughout. Regulators and platforms should test whether existing safeguards can address agent-amplified disinformation, harmful content, automated abuse and coordinated manipulation. Financial regulators and institutions should set controls on transaction permissions, volume, velocity, counterparties, authentication, anomaly detection and emergency suspension before agentic systems become deeply embedded in financial activity.

Critical-infrastructure regulators and operators should require stress testing, bounded autonomy, fallback procedures, human escalation, cyber resilience and clear allocations of responsibility among operators, vendors, developers and integrators. Governments should test the common incident protocol against failures spanning multiple sectors or regulatory regimes. Providers whose systems create shared dependencies should participate in these exercises and supply the evidence needed for a coordinated response. ASEAN bodies and other regional actors can support information sharing, capacity building, shared norms, scenario exercises and coordinated responses to cross-border risks. Academia and civil society should research how risks propagate across sectors and borders, contribute to scenario exercises and evaluate effects on public trust, labour, information ecosystems and regional resilience.

This recommendation is necessarily preventive. Configuration standards, transaction safeguards, regional protocols and incident playbooks are most valuable when developed before agentic systems become deeply embedded and dependencies are difficult to unwind. Exercises should test not only whether an individual organisation can recover, but whether authorities and sectors can recognise a

shared incident, exchange evidence and coordinate intervention quickly enough to prevent cascading harm.

Implementation at a Glance

The table below provides a clearer overview of what needs to be implemented, by whom, and where implementation could begin:

Recommendation	Singapore mechanism	Lead actors	Supporting actors	First steps
<i>1. Agent identity, permissioning and accountability</i>	Agent identity pilots and sector playbooks	Enterprises; sector regulators; providers and platforms	Financial institutions; AI governance professionals	Inventory agents and owners; classify risk; run an identity pilot in a high-trust setting
<i>2. Use-case-specific human oversight</i>	Sector oversight playbooks and assurance sandboxes	Deploying enterprises; sector regulators; providers and platforms	Domain experts; employees; academia	Match common use cases to oversight models; test whether people can understand, interrupt and recover from agentic action
<i>3. Liability and incident response</i>	Common incident protocol with sector annexes	Governments and regulators	Legal professionals and academics; enterprises; providers; platforms; insurers	Map applicable law; allocate contractual duties; issue and test a common incident protocol
<i>4. Organisational transformation</i>	Procurement expectations and shared sandboxes	Enterprise leadership and boards; public procurers	Governance teams; HR; employees; industry associations; public agencies; academia; providers and platforms	Set risk appetite; adopt procurement expectations; expand sandbox access; map workforce effects
<i>5. Systemic and regional resilience</i>	Cross-sector and regional incident exercises	Governments; cybersecurity, financial and infrastructure regulators; platforms	Enterprises; ASEAN bodies; law enforcement; academia and civil society	Identify shared dependencies and shadow agents; run cross-sector exercises; establish information-sharing pathways

Conclusion: A Shared Mandate for Responsible Agentic AI Deployment

The roundtable began with a practical question: how can enterprises move from experimentation with agentic AI towards responsible deployment? Its answer was that neither principles nor technical capability is sufficient. Governance must be operational, risk-sensitive and embedded in the systems, organisations and institutions through which agents act. It must shape permissions, monitoring, human intervention and accountability at the point of action. The gaps are interdependent. Identity and permissioning make activity visible; use-case-specific oversight preserves meaningful human judgement; clearer liability and incident arrangements make responsibility legible; organisational capacity and workforce participation make governance function in practice; and cross-sector coordination addresses failures that extend beyond individual enterprises. No single registry, oversight rule, liability regime or technical safeguard can do all of this.

Responsibility must be shared without becoming diffuse. Regulators should turn the operational mechanisms proposed in this report into sector-specific expectations and coordinate across institutional boundaries. Enterprises should define risk appetite, involve employees and domain experts and establish controls before deployment scales. Providers and platforms should supply the common baseline on which effective enterprise governance depends. Industry associations, academia, civil society and regional bodies should support smaller organisations, scrutinise wider effects and prepare for cross-border risks. The objective is proportionate governance. A low-risk internal productivity agent should not be governed like a system that conducts financial transactions, affects legal rights or interacts with critical infrastructure. Calibrating controls to what a system can do and the consequence of failure makes governance part of the infrastructure for responsible adoption—enabling organisations to move beyond experimentation while preserving accountability, trust, human capability and public resilience.

Funding: This roundtable was funded via a charitable grant from [Google.org](https://www.google.org/), as part of Google’s [Digital Futures Project](#).

