WEP Warwick Evans Publishing

Academic Journal of Management Science and Engineering

ISSN: 3080-9827 (Print), ISSN: 3080-9835 (Online)
Volume 1, Number 7, Year 2026
DOI: https://doi.org/10.62051/ajmse.v1n7.01



Cybersecurity Responsibilities in Enterprise Generative AI Adoption: A Qualitative Documentary Analysis of Delegated Authority and Organizational Oversight

Xiaohang Li [1], Chengzhi Cai [2], Haizhou Wang [3, *]

[1] School of Economics and Management, Beihua University, Jilin, Jilin, 132013, China
[2] The High School Affiliated to Yunnan Normal University, Kunming, Yunnan, 650106, China
[3] Apex Nexus Pte Ltd, Singapore, 159471, Singapore
* Corresponding Author: Haizhou Wang (Email: haizhouwang2025@gmail.com)

ABSTRACT

Enterprise generative artificial intelligence assistants can retrieve organizational information, recommend decisions, and initiate actions through connected tools. These capabilities make cybersecurity a question of how authority and responsibility are distributed across systems and organizations. This study examines how public guidance translates security responsibilities into controls over enterprise AI assistants. A purposive corpus of 12 documentary units from public bodies, a security community, and technology providers was retrieved on 28 September 2026. Focused qualitative document analysis combined six initial coding categories with cross-document comparison, producing a register of 27 evidence records. Four interpretive themes were developed: responsibility is allocated at different organizational and technical levels; permission inheritance requires contextual judgment about appropriate access; human oversight depends on enforceable intervention points; and incident accountability requires continuity across organizational and temporal boundaries. Counterevidence matters: the corpus contains explicit oversight, authorization, and incident-management provisions, so it does not support a general claim that AI security guidance lacks responsibility requirements. The analysis instead identifies an implementation translation problem between broad obligations, application-level controls, and locally assigned decision rights. A proposed governance framework links retrieval, recommendation, and action capabilities to permission boundaries, intervention conditions, accountable owners, and retained decision evidence. The framework is an analytical synthesis, not an empirically validated maturity model. The study contributes a focused account of responsibility at transitions between AI capabilities and offers an auditable structure for enterprise deployment decisions. Its conclusions concern the selected documents and do not establish organizational compliance, control effectiveness, or industry-wide prevalence.

KEYWORDS

Generative Artificial Intelligence; Cybersecurity Governance; Enterprise AI Assistants; Delegated Authority; Human Oversight; Qualitative Document Analysis.

1. INTRODUCTION

The deployment of an enterprise AI assistant is also a decision about authority. An assistant used to locate a policy document operates within a different decision setting from one that recommends a security response or changes a user's access permissions. Yet these capabilities may appear within the same conversational interface. A familiar interface can therefore conceal a substantial change in

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

what the surrounding system is permitted to do. For cybersecurity governance, the important question is not simply whether the model produces convincing answers. It is whether an organization can explain which resources the assistant may use, which consequences it may initiate, and who can intervene when a proposed operation exceeds its intended purpose.

Research on indirect prompt injection demonstrates why the surrounding application matters. Greshake et al. (2023) show how instructions embedded in retrieved content can influence an LLM-integrated application without the attacker directly controlling the user's prompt. The security problem consequently extends beyond a model's response to a cooperative user. It includes the relationships between the user, externally supplied information, connected tools, and downstream systems. AgentDojo similarly examines agents that act through tools while encountering untrusted data, separating successful completion of a legitimate task from resistance to an attack (Debenedetti et al., 2024). Neither research program, by itself, determines how a deploying organization should allocate approval and response responsibilities.

Public guidance offers several points of entry. The NIST AI Risk Management Framework identifies organizational responsibilities, while its generative AI profile elaborates actions for specific risk contexts (NIST, 2023, 2024). NCSC guidance addresses security across the system lifecycle, and the UK AI Cyber Security Code of Practice assigns provisions to stakeholder groups (NCSC, 2023; DSIT, 2025). Technical risk guidance and provider documentation operate at other levels: an application vulnerability, a permission model, or a service boundary. Their coexistence raises a practical analytical question. How does an enterprise connect obligations expressed at these different levels when approving a particular AI-enabled workflow?

This article studies that connection through a bounded qualitative analysis of public documents. Its object is the representation of cybersecurity responsibility, not the incidence of cyberattacks or the behavior of employees. A responsibility representation identifies or implies an actor, an expected action, and a domain over which that actor is expected to exercise control. Some representations are general, such as an obligation to manage risk. Others identify a concrete point at which a request should be checked or an incident should be escalated. Comparing them makes it possible to examine how organizational accountability relates to the technical ability to prevent, approve, detect, or reverse an operation.

Three research questions guide the analysis. First, how do the selected documents allocate cybersecurity responsibilities among providers, deploying organizations, and users? Second, how do they connect those responsibilities to data access, output verification, and AI-assisted actions? Third, what implementation decisions remain when provisions expressed at different levels are brought together? The third question deliberately avoids treating contextual flexibility as evidence of failure. A general framework cannot identify the appropriate approver for every company, and a short product page cannot reasonably substitute for an enterprise security policy.

The contribution is therefore narrower than a claim to discover an entirely new AI risk. The study develops a cross-level explanation of how existing responsibilities must be translated at the point where a capability is delegated. It distinguishes the authority to retrieve information, the authority to influence a decision, and the authority to execute an operation. It then proposes a record structure that connects these capabilities to local decision rights and evidence. This makes the contribution relevant to information systems and cybersecurity management research, while retaining a direct connection to the technical characteristics of tool-using AI applications.

The findings challenge a simple account of an accountability vacuum. Several sources already specify substantive safeguards and responsibility allocations. The interpretive result is that the implementation problem lies in joining those provisions into a workflow that an enterprise can operate and examine. The remainder of the article establishes the conceptual approach, describes the corpus and coding procedure, presents four themes, and discusses the proposed framework and its limits.

2. LITERATURE AND CONCEPTUAL APPROACH

2.1. From Model Behavior to Application Authority

An enterprise assistant is treated here as an application that uses generative AI to help users work with organizational information or systems. It may be a hosted productivity tool or an application assembled from models, retrieval components, and service integrations. This definition does not equate every assistant with an autonomous agent. The relevant distinction is the capability delegated in a particular workflow, rather than the product label. A system may remain conversational while obtaining permission to perform consequential operations. Conversely, a technically sophisticated agent may be confined to a narrow and reversible task.

The prompt-injection literature makes the application boundary especially important. The distinction between a trusted instruction and information encountered during a task is difficult to preserve when both are interpreted by a language model (Greshake et al., 2023). However, susceptibility at the model level does not determine every possible consequence. Consequences depend on the permissions and interfaces available to the wider system. OWASP separates prompt injection, sensitive information disclosure, improper output handling, and excessive agency into distinct entries (OWASP, 2025a–d). These categories are useful for this study because they direct attention to different points at which responsibility can be exercised rather than suggesting that all risks have one cause.

Recent technical work also complicates the assumption that human checking is the only available response. CaMeL introduces a protective layer that separates control and data flows and uses capabilities to constrain information movement (Debenedetti et al., 2025). The relevance here is conceptual: control can be enforced outside the model, even when the model itself remains fallible. The study does not evaluate that architecture or generalize its benchmark results to enterprise deployments. Instead, it asks what organizational decisions are needed to specify the boundaries that such technical mechanisms would enforce.

2.2. Accountability within a Sociotechnical System

A sociotechnical perspective treats the system boundary as including people, organizational practices, and technical components. Selbst et al. (2019), writing about fairness, explain how abstracting a technical system away from its social setting can distort the problem being addressed. Their argument is not evidence about enterprise AI cybersecurity. It provides a methodological warning against reducing accountability to a setting inside a model or a checklist attached to a product. In the present context, the authority to approve an action may belong to a manager, the technical restriction may be administered by an IT team, and the evidence required for investigation may be retained by a provider.

This study distinguishes responsibility, authority, and accountability. Responsibility refers to an assigned duty to perform or maintain a control. Authority refers to the power to permit, block, or change an operation. Accountability refers to the ability to explain and justify the resulting decisions to an appropriate reviewing party. These definitions are working analytical distinctions rather than a legal classification. They allow the analysis to identify arrangements in which a person is responsible for reviewing outputs but lacks relevant information, or a technical administrator can alter permissions without owning the business consequences of doing so.

Raji et al. (2020) connect organizational accountability to an end-to-end internal auditing process and the documents produced along that process. Their framework supports the importance of evidence continuity across development and deployment. The present study concentrates on a smaller object: a delegated capability within an enterprise workflow. The proposed contribution is not another complete auditing methodology. It is a way of connecting an authorization decision to its implementation and subsequent review, so that evidence about those stages can be interpreted together.

2.3. Human Oversight and Appropriate Reliance

The presence of a human does not establish the quality of oversight. A user may see an AI recommendation but have little basis for challenging it. A reviewer may possess expertise but receive the information only after an irreversible operation. An administrator may be able to stop a system but lack responsibility for monitoring it. These are analytically different weaknesses, although each can occur in a workflow described as involving human oversight.

Experimental work by Buçinca et al. (2021) found that interventions encouraging more deliberative engagement reduced overreliance compared with simpler explanation approaches in their study, while also creating user-experience trade-offs. This does not establish that the same intervention would be effective for enterprise cybersecurity. It does establish a reason to avoid using the availability of explanations as a proxy for effective human judgment. Oversight should be examined in terms of the decision a person can make and the conditions under which that decision is made.

For this article, oversight is operationally meaningful when a reviewer can access relevant evidence, has a defined intervention opportunity, and can affect the outcome. These conditions are proposed as an analytical lens, not reported as a validated measurement scale. They accommodate both manual approval and technically enforced limits. A narrowly specified automated check may be more appropriate for a routine operation than a confirmation prompt that a user cannot evaluate. A consequential exception may require a person with business authority, even when the technical validation succeeds.

2.4. Delegated Capability as the Unit of Interpretation

Three capability configurations guide comparison. Retrieval involves obtaining information from approved sources. Recommendation involves generating material intended to influence a human decision. Action involves initiating a change in an external system. The categories can overlap, and they are not a universal ordering of risk. Retrieving highly confidential material may have more serious consequences than executing a reversible update to a low-value record. The categories identify different decision boundaries, which must then be interpreted alongside sensitivity, exposure, reversibility, and possible impact.

The analytical question is whether the controls and responsibilities described for a workflow remain appropriate when one configuration changes into another. The same assistant might first summarize a security alert, then recommend account suspension, and finally suspend an account through an integration. Examining the transitions makes visible several separate decisions: who may see the alert, who judges the recommendation, and who authorizes the suspension. This focus avoids treating permission to use an assistant as blanket permission for every operation the assistant can technically perform.

3. METHODOLOGY

3.1. Research Design and Scope

The study uses focused qualitative document analysis. Documents are treated as evidence of how their issuers articulate responsibilities and controls, consistent with the use of documentary material as a qualitative data source discussed by Bowen (2009). The analysis is interpretive and comparative. It is not a systematic review of all AI security guidance, a survey of organizational practices, or a compliance assessment. The corpus was chosen to compare different levels of responsibility specification, rather than to estimate how frequently particular practices occur.

A matrix-based approach was used to organize evidence by source and analytical category, drawing on the structured comparison described by Gale et al. (2013). The procedure combined an initial set

of categories with revisions arising from the documents. It does not claim grounded-theory saturation, independent double coding, or inter-rater agreement. The analysis and manuscript were prepared with generative AI assistance; no independent human coding validation is claimed. These conditions limit the evidential status of the interpretation and are considered explicitly in Section 6.

3.2. Corpus Construction

The final corpus contains 12 documentary units retrieved on 28 September 2026. Selection was purposive and used three inclusion conditions: public availability in English; identifiable institutional or organizational authorship; and substantive material addressing responsibility or security controls relevant to enterprise AI applications. General AI frameworks were included when their provisions applied to deployment and organizational oversight. Sources that only advertised capabilities, repeated another page without adding an analytical role, or reported model performance without governance content were not used as primary documentary units.

Retrieval used targeted web searches, direct visits to issuing organizations, and navigation from relevant sources. Search phrases included "generative AI profile cybersecurity human AI configuration," "Guidelines for secure AI system development," "Microsoft Copilot data privacy security permissions," and "Amazon Bedrock shared responsibility security." This was a targeted discovery process rather than an exhaustive database search. No numerical screening flow or total number of excluded search results is claimed. Supporting academic literature was retrieved separately and was not included in the documentary corpus.

Table 1. Primary documentary corpus

ID	Document and source	Version or date	Genre and analytical focus
D01	NIST AI RMF 1.0 Source	2023	Governance framework; GOVERN 2 and 3
D02	NIST Generative AI Profile Source	2024	Risk guidance; inventory and incident actions
D03	NCSC secure AI guidelines Source	2023, v1.0	Lifecycle guidance; four linked chapters
D04	UK AI Cyber Security Code Source	2025	Voluntary code; roles and control provisions
D05	OWASP Prompt Injection Source	2025 edition	Technical risk entry; untrusted inputs and approval
D06	OWASP Sensitive Information Disclosure Source	2025 edition	Technical risk entry; data boundaries
D07	OWASP Improper Output Handling Source	2025 edition	Technical risk entry; downstream validation
D08	OWASP Excessive Agency Source	2025 edition	Technical risk entry; permissions and execution
D09	Microsoft Copilot privacy and security Source	Live; updated 18 Aug 2026	Product documentation; access and retention
D10	Amazon Bedrock security Source	Live; accessed 28 Sep 2026	Service documentation; shared responsibility
D11	Amazon Bedrock incident response Source	Live; accessed 28 Sep 2026	Service documentation; response boundary
D12	Google SAIF introduction Source	8 Jun 2023	Conceptual framework; detection and business context

The corpus comprises two NIST publications, one NCSC guideline collection, the UK voluntary code, four OWASP risk entries, one Microsoft product document, two Amazon Bedrock documentation pages, and Google's introductory SAIF document. NCSC's four lifecycle chapters are treated as one document family, rather than four independent observations. The four OWASP entries are separate analytical units but belong to one taxonomy. The AWS pages similarly share a provider and

responsibility model. These dependencies mean that agreement across documents cannot be interpreted as agreement among 12 independent institutions.

The original research concept contemplated a larger collection. For this bounded analysis, 12 units were retained to permit direct comparison with identifiable locators. This is a practical scope decision, not a saturation claim. The corpus includes foundational documents issued from 2023 onward and live documentation as accessed in September 2026. It is not a historical reconstruction of what all sources stated at a common earlier date. Table 1 records the corpus; Appendix A identifies the selected provisions used in the analysis.

3.3. Extraction and Coding

The unit of extraction was a bounded passage describing a responsibility, permission condition, security measure, or response arrangement. A passage could comprise a numbered provision or several closely related sentences under a heading. The analysis retained 27 evidence records. This number describes the register supporting the article, not the total number of potentially relevant statements in the documents. Each record contains a source identifier, a stable section or action locator, a concise paraphrase, and one or more codes. Paraphrases are analytical summaries and are not presented as quotations.

Six categories were specified at the outset: data access, output verification, action authorization, responsibility allocation, incident management, and control specificity. Table 2 gives their operational meanings and boundaries. Coding permitted overlap. For example, a requirement to check permissions before executing a tool call can concern both access and action authorization. A statement recommending logging belongs to incident management when it supports investigation; it is not automatically evidence of an approval control. This separation prevents monitoring after an event from being treated as equivalent to preventing the event.

Table 2. Initial coding framework

Code	Category	Included meaning	Boundary rule
DA	Data access	Permissions, sources, identities, and sensitive information	Exclude output accuracy without an access issue.
OV	Output verification	Evaluation or validation before reliance or downstream use	Distinguish human judgment from code-level validation.
AA	Action authorization	Permission to initiate or approve an external operation	Do not equate permission to view with permission to act.
RA	Responsibility allocation	Named actor, role, organizational boundary, or duty	Do not infer a local job title from a broad actor group.
IM	Incident management	Detection, investigation, response, recovery, or supporting records	Separate retrospective evidence from prior prevention.
CS	Control specificity	Explicit trigger, enforcement point, record, or reassessment condition	Descriptive code; not a numerical quality rating.

Coding proceeded through three passes. The first identified explicit actors and control objects in the selected passages. The second compared the trigger for a control, where it could be enforced, and whether a record or communication was specified. The third compared initial interpretations against counterexamples. This last pass was consequential: explicit provisions for oversight, downstream authorization, and third-party incident planning prevented the analysis from characterizing the corpus as generally silent about responsibility. The resulting themes describe relationships among provisions, rather than frequencies of keywords.

Four interpretive labels emerged during comparison: different levels of responsibility allocation, context-dependent access decisions, enforceable intervention points, and continuity of incident evidence. These labels organize the findings but do not replace the original source wording. In the evidence register, initial codes remain visible so that a reader can distinguish a document-level observation from the broader interpretation developed from it. Section 5 separately presents recommendations derived from the comparison; these recommendations are not attributed to the source organizations.

3.4. Evidence Quality and Interpretive Safeguards

The analysis distinguishes three evidential levels. Documentary observations report what a selected passage says. Interpretations explain relationships between such observations. Proposals specify arrangements that an enterprise could adopt. Keeping these levels separate is important because a recommendation in a public document does not demonstrate implementation, and a plausible implementation design does not establish effectiveness. Provider documentation is treated as a public statement of service arrangements, not independently tested evidence that the service always behaves as described.

Negative-case checking focused on passages that challenged the starting concern about ambiguous oversight. NIST's role provisions, the UK code's oversight requirements, and OWASP's authorization guidance provided such counterevidence. Document genre was also treated as an alternative explanation. A broad framework leaves local names unspecified for a different reason from a product page describing inherited permissions. Neither omission is sufficient to establish negligence, incompleteness of an entire framework, or a defect in the service concerned.

Source URLs, retrieval dates, and provision locators support inspection. Local retrieval copies were used where available; the two short AWS pages were examined through their retrieved web text because the direct download returned a minimal page. The study does not claim a publicly deposited archive or a complete version history. For live pages, the access date defines the analyzed snapshot. The research involved no interviews, personal records, organizational access, or interaction with operational AI systems. No institutional ethics approval or exemption determination is claimed.

4. FINDINGS

4.1. Responsibility is Allocated at Different Levels

The first theme is a difference in the level at which responsibility is expressed. NIST's framework assigns organizational duties for roles, communication, training, and leadership decisions (D01; E01–E02). The UK code identifies stakeholder groups that may occupy overlapping positions in a supply chain (D04; E09). AWS documentation separates the provider's infrastructure role from the customer's responsibilities within the service context (D10; E24). These are identifiable allocations, rather than evidence that no actor is responsible. However, they answer different questions about the location of responsibility.

The cross-document interpretation is that role clarity at one level does not automatically produce an executable arrangement at another. A provider–customer distinction says which organization must address a class of issue. It does not identify which person inside the customer organization may approve an unusual disclosure. An internal role policy can identify a business owner but still leave an integration using an overly broad service identity. Both allocations may be clearly written while the relationship between them remains a local implementation task.

An enterprise consequently needs to distinguish an organizational owner from the actor implementing a control. A business process owner may accept the consequences of a workflow, while an identity administrator configures its service permissions. A reviewer may decide whether one particular action

should proceed. Treating all three as a single “responsible person” would obscure the different forms of authority involved. This is an interpretive implication of comparing the sources, not a finding that the organizations producing them have made that mistake.

The strongest counterexample to an accountability-vacuum interpretation is the specificity already present in the governance material. NIST differentiates human–AI responsibilities, and the UK code links oversight to technical measures where it functions as a control (D01, D04; E02, E11). The appropriate conclusion is therefore conditional. Public guidance supplies substantive components of an allocation structure, but a deploying enterprise must connect those components to its own personnel, configurations, and approval processes. The study cannot determine whether enterprises routinely make that connection because it contains no evidence from their internal operations.

4.2. Permission Inheritance Requires a Judgment About Appropriate Access

The second theme concerns the distinction between enforcing an existing permission and deciding that the permission is appropriate for the task. Microsoft's documentation states that Copilot respects users' existing access and identifies administrative controls over agents (D09; E21–E22). OWASP's disclosure guidance emphasizes necessary access and restricted data sources (D06; E15). The comparison draws attention to the organizational decision upstream of technical enforcement: which information should be accessible to this user, through this workflow, for this purpose?

Permission inheritance is meaningful counterevidence to the assumption that introducing an assistant necessarily bypasses access controls. The source describes a boundary that the application is intended to preserve. The interpretive issue is different. An inherited entitlement may be broader than the information needed for a particular task. A system can therefore conform to an existing access rule while prompting an organization to reconsider whether that rule still expresses its intended information-sharing policy. This does not establish a vulnerability in the provider's implementation or prove that inappropriate exposure has occurred.

The distinction can be illustrated without inventing a case study. Consider a hypothetical employee who legitimately has access to several project repositories but needs only one repository for a current assignment. The research question is not whether the employee's general access is unlawful or technically invalid. It is whether an assistant's retrieval context should use every available repository, a selected subset, or an explicitly approved collection. Those alternatives represent different governance choices. Their suitability depends on the intended use, confidentiality, and the consequences of combining information across contexts.

The documents also differentiate content handling from human behavior. NCSC's development guidance treats generated material and logs as assets requiring protection appropriate to their sensitivity (D03; E06). This broadens the object of control beyond the original repository. A response, cached context, or audit record may reproduce information that was properly accessed at the source. The analytical implication is that responsibility should follow the information into the artifact created by the assistant. Authorization to read a source does not, by itself, settle the conditions for sharing or retaining a resulting synthesis.

The theme is therefore not that inherited permissions are inadequate in general. It is that permission enforcement and contextual access governance are complementary tasks. The former can be implemented consistently by a platform; the latter requires a decision about the purpose of the enterprise workflow. A useful assessment must ask both whether the intended rule is enforced and whether the rule still represents the organization's intended boundary.

4.3. Human Oversight Depends on Enforceable Intervention Points

The third theme is the distinction between evaluating an answer and authorizing an operation. OWASP's improper-output-handling entry concerns validation before model output reaches other

components, whereas its excessive-agency entry addresses the functions, privileges, and autonomy available to an application (D07–D08; E16–E20). Prompt-injection guidance also identifies human approval for privileged operations (D05; E14). Together, these provisions show that oversight is already linked to technical control points in the corpus.

The interpretation developed here is that different forms of review should not be substituted for one another. A reviewer may find a recommendation factually persuasive without being authorized to approve its execution. Conversely, an authorized decision-maker may approve a legitimate business objective while the proposed technical operation still contains unsafe parameters. Content review, business authorization, and technical validation answer separate questions. A workflow can require all three, although the same person or component may perform more than one function in a simple setting.

OWASP provides particularly important counterevidence to a reading that would place all confidence in a model's self-restraint: its excessive-agency guidance places authorization checks downstream and limits permissions and high-impact actions (D08; E18–E20). The interpretive significance is that security responsibility can be embodied in an enforcement point rather than expressed only in instructions to an assistant. This is compatible with permitting a model to generate flexible proposals while restricting what the surrounding system will actually execute.

A hypothetical account-suspension workflow clarifies the distinction. The assistant might summarize evidence, recommend suspension, and prepare an operation. A technically valid request could still target the wrong account or conflict with business continuity requirements. An approval should therefore concern the specified account and operation, not merely the general idea of "following the recommendation." This example is an analytical illustration. No account-management system was tested, and no conclusion about the success rate of such a control is drawn.

Oversight also has a timing dimension. A review conducted after execution supports accountability and learning but cannot prevent the completed operation. For reversible actions, subsequent review may be proportionate; for actions with lasting consequences, a prior decision point may be necessary. The corpus supports distinguishing these control functions. It does not supply a universal rule for the threshold between them. That threshold remains a contextual decision involving consequences, reversibility, available evidence, and the feasibility of a meaningful review.

The theme revises the original concern about vague human supervision. The more defensible finding is that relevant intervention mechanisms exist in public guidance, but enterprises must decide where to place them in a workflow and who can exercise them. An approval interface without a defined decision right is not equivalent to authorization. Equally, a written decision right without an enforceable interruption point cannot ensure that the operation will wait for that decision.

4.4. Incident Accountability Requires Continuity Across Boundaries

The fourth theme concerns the evidence needed after a failure or suspected compromise. NIST's generative AI profile addresses ownership and rehearsal of third-party incident response (D02; E04). AWS's incident page distinguishes service-side response from customer-side response (D11; E25). NCSC calls for incident procedures and supporting logs, while Google's SAIF introduction connects detection and response with wider security functions (D03, D12; E07, E26). These sources do not support a blanket claim that incident ownership has been ignored.

The comparative interpretation is that a division of responsibilities creates a coordination requirement. A provider can investigate the service it operates, while a customer investigates the workflow and permissions it configured. A business owner may need to determine which consequences occurred outside the service. Reconstructing one event can require evidence from each of these locations. The practical problem is therefore not only naming a responder. It is ensuring that the responders can relate their records to the same action, configuration, and decision.

Evidence continuity is also temporal. An action may be approved under one configuration and executed after a connector, prompt, or permission has changed. NCSC's maintenance guidance treats changes to system components as potentially consequential for behavior, and the UK code includes reassessment when settings change (D03–D04; E08, E12). These observations support examining authorization as a decision about a particular configuration, rather than an indefinite approval of a product name. They do not show that any observed enterprise has failed to reassess a change.

Logging introduces its own governance choice. Rich records can assist investigation but may reproduce confidential prompts, outputs, or business context. Microsoft's documentation describes retention management for interaction data, and NCSC treats logs as sensitive assets (D09, D03; E23, E06). The implication is that maximum collection is not automatically the best accountability arrangement. An enterprise needs sufficient evidence to reconstruct an event, together with a decision about who may access that evidence and how long it is retained.

The four themes are summarized in Table 3. Their common thread is the need to connect allocations expressed at different levels. They should not be interpreted as rankings of the source organizations or measurements of document quality. The sources serve different audiences and functions. The analytical contribution lies in showing the decisions that arise when they are used together for an enterprise workflow.

Table 3. Cross-document interpretations and counterevidence

Theme	Evidence anchors	Interpretation	Important qualification
Allocation level	D01, D04, D10	Connect organizational, supply-chain, and service responsibilities.	Explicit role provisions already exist.
Access context	D03, D06, D09	Distinguish enforced entitlements from task-specific access decisions.	Permission inheritance is a substantive boundary.
Intervention	D05, D07, D08	Separate content review, business approval, and execution checks.	Technical authorization is explicitly addressed.
Evidence continuity	D02, D03, D09, D11, D12	Connect response evidence across configurations and organizations.	Incident duties are not generally absent.

5. DISCUSSION

5.1. An Implementation Translation Problem

The findings suggest an implementation translation problem rather than a universal absence of accountability guidance. A general obligation to maintain oversight, a technical rule requiring authorization, and a service statement about customer responsibility may all be appropriate. The remaining work is to connect them to a concrete operation. This translation is where an enterprise specifies which business role owns the consequences, which technical boundary limits the operation, and which evidence supports approval or subsequent explanation.

The proposed concept of capability–responsibility alignment describes that connection. Alignment exists, in the analytical sense used here, when a delegated capability has a defined scope, an actor with appropriate decision rights, an intervention mechanism, and a record sufficient to explain its use. This is not a binary certification criterion. An enterprise may implement it in several ways, and a complete written arrangement may still fail in practice. The concept identifies questions that can be investigated; it does not replace testing, professional judgment, or organizational inquiry.

This account adds specificity to sociotechnical reasoning by locating the relevant boundary at a transition in authority. Moving from retrieval to recommendation changes the relationship between information and judgment. Moving from recommendation to action changes the relationship between judgment and execution. A study focused only on model output may miss the second transition. A

study focused only on identity permissions may miss whether the person approving execution understood the proposed action. Examining both makes the organization and the application part of the same analytical object.

The argument also clarifies what would count as a genuine gap. A gap is not established because a public framework leaves an employee's name unspecified. More convincing evidence would show that an enterprise has delegated an operation while no actor can explain its permissible scope, approve exceptions, or reconstruct its execution. Such evidence requires fieldwork or system examination. The present documents make that proposition researchable but do not establish its prevalence. This distinction is essential to avoid turning a conceptual concern into an unsupported empirical finding.

5.2. A Framework for Governing Delegated Capabilities

Table 4 translates the themes into a proposed framework. It treats retrieval, recommendation, and action as capability configurations, not maturity stages. Each configuration is assessed against four elements: its permitted boundary, the condition requiring intervention, the accountable organizational role, and the evidence retained. The elements are intentionally few so that the framework can attach to an existing deployment or change process instead of becoming a separate AI bureaucracy.

Table 4. Proposed governance framework for delegated capabilities

Capability	Boundary and intervention	Accountable role to assign locally	Evidence to retain
Retrieval	Approved sources and identity scope; review new sources or audiences.	Information owner with the access-control administrator.	Source scope, permission basis, configuration reference.
Recommendation	Defined decision purpose; review consequential reliance or uncertainty.	Process owner and a reviewer competent for the decision.	Recommendation, verification basis, accepted or revised decision.
Action	Allowed operations and targets; approve exceptions before execution where required.	Business authority supported by technical control ownership.	Specific authorization, policy check, execution result, recovery route.

Note. This is an analytical proposal, not a tested control standard or a risk ranking. Sensitivity, consequences, and reversibility modify all three configurations.

For retrieval, the primary decision concerns the information context. A deployment record should identify approved collections and the basis for access, including whether the assistant uses a user's identity or a service identity. Intervention might be triggered by a request involving a new repository or a different audience for the resulting output. Relevant evidence would show the source boundary and the applicable permission configuration. This proposal does not require recording all document contents; references and configuration information may be sufficient for many accountability purposes.

For recommendation, the decision concerns reliance. The enterprise should identify what type of judgment the output is intended to support and when that judgment requires independent verification. A recommendation used to draft an internal note differs from one used to restrict a customer's account. The responsible role should have expertise relevant to the consequence, rather than merely familiarity with the AI interface. Evidence should distinguish the model's proposal from the decision that a person actually made. This preserves the possibility of identifying a justified rejection or modification as a successful use of the assistant.

For action, the decision concerns execution within a permitted operating range. The organization should identify allowed operations, protected targets, and circumstances requiring escalation. Approval, where required, should bind to a sufficiently specific operation and remain subject to technical authorization. An execution record should connect the request to the approved scope and resulting outcome. For some routine and reversible operations, predefined rules may be appropriate.

For consequential exceptions, a designated person may be necessary. The framework does not assume that every action requires a separate human confirmation.

Across all three configurations, a material change should reopen the relevant decision. Adding a new connector might alter the retrieval boundary without changing the language model. Allowing write access might change the action boundary while leaving the interface untouched. Changing the intended recipients of an output might alter confidentiality exposure without changing either the model or the source repository. A change process centered only on model versions would miss these possibilities. The proposed framework instead asks which previously accepted conditions no longer hold.

5.3. Implications for Oversight and Audit Evidence

The findings support treating oversight as a designed responsibility with an observable decision point. This does not mean requiring a large amount of documentation for every use. It means retaining enough information to determine what was delegated and whether the actual operation remained within that scope. A compact record can identify the workflow, owner, approved capability, intervention conditions, relevant configuration, and review date. Its value depends on accuracy and connection to the deployed system, not on length or the sophistication of the template.

The relationship to internal auditing is complementary. Raji et al. (2020) emphasize a documentary process across AI development. The proposed record provides one object that such a process could examine. An auditor could compare the stated capability with actual permissions and compare approval requirements with the operations that occurred. That examination might reveal a mismatch, but the record alone would not prove one. Documentary completeness and operational assurance should remain separate judgments.

An organization should also resist equating more manual review with better control. Human review consumes attention and can become routine acceptance when the person lacks context or meaningful options. The appropriate response may be to narrow the action space, improve the evidence presented, or remove a capability whose benefits do not justify its oversight burden. Buçinca et al. (2021) provide a reason to take the design of reliance seriously, while the present framework specifies the organizational question that such design should answer: what decision is this person being asked to make, and can that decision influence the outcome?

The evidence required for an investigation should likewise be proportionate. Reproducing every prompt and response in a widely accessible log could create unnecessary exposure. Conversely, retaining only a generic statement that an assistant was used may prevent meaningful reconstruction. The proposed balance is to identify the minimum record needed for the intended accountability purpose, then protect it according to its content. Determining that minimum requires consideration of the workflow; it cannot be inferred solely from a vendor's default logging settings.

5.4. Propositions for Subsequent Empirical Research

Three propositions follow from the synthesis and are offered for testing, not presented as confirmed results. First, deployments that link a capability-specific owner to an enforceable permission boundary should exhibit fewer unresolved responsibility disputes than deployments that rely on general acceptable-use policies alone. A comparative case study could examine decision records and interview process owners, security staff, and users. Evidence against the proposition would include clear and consistently applied responsibility arrangements that function without a dedicated capability record.

Second, oversight should be more consequential when approvals refer to a specific proposed operation and relevant context than when they merely confirm generic use of an assistant. This proposition could be examined through controlled workflow studies measuring detection of

inappropriate actions, review burden, and unnecessary rejection. The outcome should not be restricted to approval speed. A faster decision is not necessarily a better security decision, and a high rejection rate can also indicate poor usability or an excessively broad task definition.

Third, organizations that can connect provider-side events, local configuration changes, and business decisions should reconstruct AI-related incidents more effectively than those retaining disconnected records. A tabletop exercise could compare the time and information required to answer a defined set of incident questions. The study would need to account for organizational size, available expertise, and differences in incident severity. These proposed investigations would move from documentary expectations to observable practices and could establish whether the framework has practical explanatory value.

6. LIMITATIONS

The first limitation is selection. The 12 units were chosen for their relevance and contrast, not through an exhaustive search of every framework, provider, or jurisdiction. English-language and accessibility requirements privilege a particular public discourse. The corpus does not represent all enterprise AI systems or all national approaches. The presence of several sources from the same document family further limits independence. No counts of codes are interpreted as population estimates, and no claim of thematic saturation is made.

Second, the evidence is normative and descriptive documentation. It establishes what selected sources prescribe or publicly describe, not what enterprises implement. Provider pages may omit arrangements documented elsewhere, and their statements were not verified through product testing or contracts. Apparent differences can arise from audience, purpose, publication date, or the limited scope of the page. The analysis therefore does not score organizations, establish legal duties, or conclude that an omitted detail is absent from an issuer's broader documentation.

Third, the method is a focused, AI-assisted synthesis without independent human coding validation. The evidence register makes the interpretation inspectable, but inspectability is not equivalent to reliability demonstrated by a separate researcher. Generative AI may favor familiar governance categories or produce an overly coherent account from heterogeneous texts. Counterexample checking and provision-level locators reduce some risks but cannot remove them. Independent re-extraction and a comparison of alternative interpretations would strengthen subsequent research.

Fourth, the materials are time-sensitive. The study combines foundational publications with live product documentation accessed on 28 September 2026. It cannot infer a historical trend from differences between those sources, and its conclusions may require revision when product architectures or guidance change. A longitudinal study would require archived versions and a defined comparison procedure. The current analysis makes no claim to preserve a complete public archive of the material.

Finally, the proposed framework has not been implemented or evaluated in an enterprise. It may be easier to apply in a well-resourced organization than in a small business with overlapping roles. It may also need extension for multi-agent systems, dynamically created tools, and workflows spanning several organizations. The three capability configurations simplify a larger design space. Their usefulness should be judged by whether they clarify decisions and support investigation, not by whether every system can be assigned to a single category.

7. CONCLUSION

This study examined how a bounded corpus of AI governance, security, and provider documents represents cybersecurity responsibility for enterprise generative AI assistants. The documents contain substantive allocations of responsibility and concrete safeguards. Their comparison does not support

a general claim that oversight or incident accountability is absent. It instead reveals an implementation task: connecting organizational roles, application permissions, intervention conditions, and evidence across a particular workflow.

The proposed framework locates that task at the capabilities delegated to an assistant. Retrieval, recommendation, and action raise different questions about access, reliance, and execution, while sensitivity and consequences determine the required safeguards. The contribution is an analytical structure for relating those questions to accountable decisions. Its effectiveness remains to be tested through organizational research. For enterprises, the practical implication is to make the permission to use AI specific enough that someone can explain what the system was authorized to do, intervene when conditions change, and reconstruct what happened afterward.

ACKNOWLEDGMENTS

Funding: This research was funded by Enerstay Sustainability Academy (Singapore) Grant Call (Call 7/2025) _SDG (Project ID CUMT-001), Singapore.

REFERENCES

[1] Amazon Web Services. (n.d.-a). Security, guardrails, and observability in Amazon Bedrock. Amazon Bedrock User Guide. Retrieved September 28, 2026, from https://docs.aws.amazon.com/bedrock/latest/userguide/security.html.

[2] Amazon Web Services. (n.d.-b). Incident response in Amazon Bedrock. Amazon Bedrock User Guide. Retrieved September 28, 2026, from https://docs.aws.amazon.com/bedrock/latest/userguide/security-incident-response.html.

[3] Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. https://doi.org/10.3316/QRJ0902027.

[4] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 188. https://doi.org/10.1145/3449287.

[5] Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems, 37. https://proceedings.nips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html.

[6] Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2025). Defeating prompt injections by design. arXiv preprint arXiv:2503.18813. https://doi.org/10.48550/arXiv.2503.18813.

[7] Department for Science, Innovation and Technology. (2025, January 31). Code of practice for the cyber security of AI. UK Government. https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai.

[8] Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Medical Research Methodology, 13, 117. https://doi.org/10.1186/1471-2288-13-117.

[9] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv preprint arXiv:2302.12173. https://doi.org/10.48550/arXiv.2302.12173.

[10] Hansen, R., & Venables, P. (2023, June 8). Introducing Google's Secure AI Framework. Google. https://blog.google/innovation-and-ai/technology/safety-security/introducing-googles-secure-ai-framework/.

[11] Microsoft. (2026, August 18). Data, privacy, and security for Microsoft Copilot. Microsoft Learn. Retrieved September 28, 2026, from https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-privacy.

[12] National Cyber Security Centre. (2023, November 27). Guidelines for secure AI system development (Version 1.0). https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

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

[14] National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1.

[15] OWASP. (2025a). LLM01:2025 Prompt injection. OWASP Gen AI Security Project. Retrieved September 28, 2026, from https://genai.owasp.org/llmrisk/llm01-prompt-injection/.

[16] OWASP. (2025b). LLM02:2025 Sensitive information disclosure. OWASP Gen AI Security Project. Retrieved September 28, 2026, from https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/.

[17] OWASP. (2025c). LLM05:2025 Improper output handling. OWASP Gen AI Security Project. Retrieved September 28, 2026, from https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/.

[18] OWASP. (2025d). LLM06:2025 Excessive agency. OWASP Gen AI Security Project. Retrieved September 28, 2026, from https://genai.owasp.org/llmrisk/llm062025-excessive-agency/.

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

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

APPENDIX A EVIDENCE REGISTER

The 27 records below are the complete retained evidence register for this focused analysis. They are concise paraphrases, not verbatim quotations, and are not a census of all provisions in the documents. Codes refer to Table 2. Section locators, rather than browser line numbers, are supplied so that the evidence can be checked on the original pages. For D03, the four lifecycle chapters form one documentary unit.

APPENDIX B RETRIEVAL AND ANALYTICAL DECISIONS

Retrieval date. All primary documentary units were consulted on 28 September 2026. The version field distinguishes dated publications from live documentation; an access date does not imply that a page was first published in 2026.

Discovery route. Initial sources were identified through targeted institutional searches for AI security guidance and followed by direct retrieval of the listed documents. Searches for enterprise permissions and shared responsibility identified the Microsoft and AWS pages. The Google SAIF introduction was included as a contrasting provider-authored conceptual framework. Literature on document analysis, sociotechnical systems, auditing, reliance, and prompt injection was used as background rather than counted as documentary evidence.

Corpus boundary. Twelve documentary units were retained. NCSC lifecycle chapters were grouped; OWASP entries were compared as related risk categories; AWS pages were treated as linked service documents. Additional pages returned during discovery were not represented as a systematically screened population. The retrieval process therefore has no PRISMA-style exclusion count.

Coding decisions. Initial categories were applied to the selected passages, followed by comparison of actor, control object, trigger, enforcement location, and evidence. Records could carry more than one code. An explicit description of inherited permissions was treated as evidence of a boundary, not evidence of unrestricted access. Post-event monitoring was distinguished from prior authorization. General requirements were not scored as defective merely because they required local implementation.

Counterexample decisions. D01 and D04 prevented a claim that human responsibility was generally undefined. D08 prevented a claim that technical guidance relied only on model instructions. D02 and D11 prevented a claim that incident ownership was generally absent.

Table A1. Retained evidence records

Record and source	Section or action locator	Paraphrased observation	Codes
E01 / D01	GOVERN 2.1–2.3	Organizational roles, training, and leadership responsibility are specified.	RA
E02 / D01	GOVERN 3.2	Human–AI oversight responsibilities are to be differentiated in policies.	RA; OV
E03 / D02	GV-1.6-003	Inventory information includes oversight roles, model versions, and access modes.	RA; CS
E04 / D02	GV-6.2-003	Third-party incident planning includes ownership, communication, rehearsal, and learning.	RA; IM; CS
E05 / D03	Secure design — Design your system for security	External-service transfers require controls appropriate to sensitive information.	DA; CS
E06 / D03	Secure development — Identify, track and protect your assets	Logs and generated content require protection reflecting their sensitivity.	DA; IM
E07 / D03	Secure deployment — Develop incident management procedures	Incident procedures and supporting audit information are expected.	IM; RA
E08 / D03	Secure operation and maintenance — Follow a secure by design approach to updates	Changes to data, models, or prompts can require renewed evaluation.	IM; CS
E09 / D04	Audience	Stakeholder categories overlap when one entity performs multiple roles.	RA
E10 / D04	Provision 2.6	Permissions on connected systems are limited by functional need and assessed risk.	DA; AA
E11 / D04	Provisions 4.1–4.3	Oversight is supported by design and technical measures where it controls risk.	OV; RA; CS
E12 / D04	Provision 3.1.1	New or updated settings trigger consideration of resulting security risks.	CS
E13 / D05	Indirect Prompt Injections	External content can introduce instructions that redirect model behavior.	DA; OV
E14 / D05	Prevention and Mitigation Strategies — item 5	Privileged operations are linked to human approval.	AA; CS
E15 / D06	Prevention and Mitigation Strategies — Access Controls	Necessary access and restricted sources constrain exposure of sensitive data.	DA
E16 / D07	Opening definition and distinction from overreliance	Downstream handling of model outputs is distinguished from broader reliance.	OV
E17 / D07	Prevention and Mitigation Strategies — items 1–5	Validation, encoding, and parameterized database operations constrain downstream interpretation.	OV; CS
E18 / D08	Prevention and Mitigation Strategies — item 4	Extension permissions should be minimized for the intended operation.	DA; AA
E19 / D08	Prevention and Mitigation Strategies — item 6	High-impact actions require approval before execution.	AA; CS
E20 / D08	Prevention and Mitigation Strategies — item 7	Downstream authorization checks enforce policy independently of model decisions.	AA; CS
E21 / D09	How does Microsoft Copilot use your proprietary organizational data	User permissions constrain organizational information surfaced by the assistant.	DA; CS
E22 / D09	Extensibility of Microsoft Copilot	Administrators control allowed agents and inspect their requested access.	DA; RA
E23 / D09	Data stored about user interactions with Microsoft Copilot	Administrative tools support retention policies for interaction data.	DA; IM; CS
E24 / D10	Security of the cloud and Security in the cloud	Provider infrastructure duties differ from customer security responsibilities.	RA
E25 / D11	Opening paragraphs	Service-side incident response is distinguished from response on the customer side.	RA; IM
E26 / D12	SAIF core element 2	AI detection and response involve monitoring and collaboration across security functions.	IM; RA
E27 / D12	SAIF core element 6	AI risks are situated within surrounding business processes.	RA; CS

D09 prevented a claim that enterprise assistants necessarily bypass existing permissions. These counterexamples changed the interpretation from an accountability vacuum to a problem of connecting provisions across levels.

Inference boundary. The framework in Table 4 and the propositions in Section 5 are original analytical proposals developed from the comparison. They are not claimed to be requirements issued by the source organizations, validated controls, or observed enterprise practices.