

DOI: https://doi.org/10.48009/2_iis_2026_104

Attribution drift in enterprise monitoring: governance, employee monitoring, and cybersecurity risk in the age of agentic AI

Johnathan Yerby, *Mercer University*, Yerby_jm@mercer.edu

Abstract

Agentic artificial intelligence is blurring the boundary between human work, automated work, and delegated digital action in enterprise environments. Enterprise monitoring systems, including endpoint detection and response, extended detection and response, security information and event management platforms, and user behavior analytics, have traditionally relied on the assumption that observed activity can be connected to a human user, service account, or system process with reasonable confidence. Agentic AI complicates this assumption because agents may operate under human credentials, through delegated permissions, as background services, or through additional accounts that are difficult for users and administrators to interpret. This paper introduces attribution drift as a conceptual construct describing the erosion of confidence that observed digital activity can be reliably attributed to intentional human conduct when semi-autonomous AI agents operate under, alongside, or near human credentials. The construct extends existing discussions of accountability gaps and responsibility gaps by focusing specifically on the monitoring, identity, and forensic conditions that precede responsibility assignment. Using a conceptual governance framework, the paper synthesizes literature on workplace monitoring, insider threat, socio-technical systems, agentic AI, and algorithmic accountability. It proposes a taxonomy of human-agent activity and develops governance propositions related to attribution confidence, employee monitoring, cybersecurity risk, and forensic readiness. The analysis argues that full AI autonomy is not required for these risks to emerge. The more immediate challenge is partial, delegated, and opaque automation within enterprise systems. Recommendations include agent identity tagging, permission separation, agent-specific audit trails, updated acceptable use policies, and incident response procedures that preserve attribution confidence while limiting unnecessary worker surveillance.

Keywords: agentic AI, employee monitoring, attribution drift, insider threat, surveillance, AI governance

Introduction

Organizations are moving from experimental use of artificial intelligence toward operational reliance on AI-assisted and AI-delegated work. Recent industry reports describe this transition through concepts such as digital labor, human-agent teams, and organizational redesign around AI agents (Microsoft, 2025). Organizations are not merely discussing AI agents as speculative future systems. They are increasingly preparing employees, workflows, and governance structures for AI systems that can perform work across enterprise applications. In some workplaces, employees are also being encouraged or expected to use AI tools even when approved technologies, permission boundaries, and monitoring expectations remain unevenly defined.

Academic literature has begun to clarify what makes agentic AI different from ordinary AI assistance. Recent surveys describe agentic AI systems as capable of pursuing complex goals, using tools, coordinating tasks, retaining memory, adapting to context, and operating with varying levels of autonomy (Acharya et al., 2025; Bandi et al., 2025; Hosseini & Seilani, 2025; Sapkota et al., 2026; Shavit et al., 2023). These capabilities matter for enterprise monitoring because they place AI systems closer to the actual performance of work. A chatbot that answers a question can usually be classified as an advisory tool. An agent that searches files, drafts communications, updates records, creates tickets, queries databases, executes scripts, or manages workflows is more difficult to classify. It may be acting for the employee, with the employee, instead of the employee, or through a technical identity that only partially corresponds to the employee.

This transition creates a governance problem for enterprise monitoring. Existing monitoring practices often depend on the ability to associate activity observed on a workstation, application, cloud service, or network with an identifiable actor. That actor might be a human user, a service account, a system daemon, a scheduled task, or a malicious process. Cybersecurity practice has never been naive enough to assume that every event is directly performed by a person. Service accounts, daemons, scheduled tasks, and enterprise security agents are not new, and mature log management practices are designed to help organizations collect, centralize, analyze, and preserve event records for security monitoring, investigations, and forensic analysis (Kent & Souppaya, 2006). In many cases, these traditional forms of automation are administratively defined, narrowly scoped, and more predictable than agentic AI systems. Agentic AI systems may instead function as semi-autonomous extensions of human users, interact across multiple tools, adapt to context, use memory, and generate activity that resembles legitimate human work (Acharya et al., 2025; Sapkota et al., 2026; Shavit et al., 2023).

The governance problem is not limited to fully autonomous AI. In some respects, partial autonomy creates greater attribution difficulty because control may shift between a human and an automated system. A digital action may be human-directed, agent-assisted, delegated, or independently executed within a permission structure that still appears to belong to the human user. As agentic AI systems evolve, organizations cannot yet rely on a consistent model for how these systems authenticate, obtain permissions, create accounts, use tokens, or operate across enterprise tools.

Consider an approved workplace AI companion that creates a background service, delegated token, or additional non-obvious or hidden account during installation or execution. Endpoint monitoring may detect the activity, but administrators may struggle to determine whether later events reflect human behavior, authorized agent behavior, benign system automation, or a security concern. This ambiguity becomes more difficult when monitoring tools themselves rely on privileged background services and non-obvious processes. In effect, one opaque automated system may detect another opaque automated system operating on a machine assigned to a human user, while the resulting activity remains difficult to meaningfully attribute.

This paper introduces attribution drift as a conceptual construct for understanding this emerging problem. Attribution drift is defined here as the gradual erosion of confidence that observed digital activity can be reliably attributed to intentional human conduct due to the increasing presence of semi-autonomous AI agents operating under, alongside, or near human credentials. Similar wording appears in other technical contexts, such as feature attribution drift in model monitoring, but the usage in this context is distinct (Amazon Web Services, n.d.). The focus here is not on changes in model explanation patterns over time. Instead, the focus is on enterprise governance, employee monitoring, insider threat analysis, forensic readiness, and accountability.

Attribution drift is related to, but distinct from, existing concepts such as accountability gaps, responsibility gaps, identity ambiguity, and attribution challenges in cybersecurity. Those concepts generally focus on the difficulty of assigning responsibility after an automated or algorithmic system produces an outcome. Attribution drift focuses on an earlier governance problem: whether the organization has enough

monitoring, identity, policy, and forensic evidence to determine what kind of actor produced the observed activity in the first place. In other words, accountability gaps ask who should be responsible for an action or harm. Attribution drift asks whether the organization can reliably distinguish direct human conduct, agent-assisted conduct, delegated agent activity, system automation, or malicious activity before responsibility is assigned.

The central argument is that agentic AI does not need to be fully autonomous to disrupt enterprise monitoring. Attribution drift emerges when human users, AI agents, credentials, service accounts, endpoint telemetry, and organizational policies become difficult to disentangle. This paper argues that attribution drift creates governance challenges for employee monitoring, insider threat analysis, surveillance practices, legal and ethical accountability, and digital forensics. To address these challenges, the paper proposes a conceptual taxonomy of human-agent activity and governance practices intended to preserve attribution confidence without defaulting to excessive employee surveillance.

Background and Literature Review

Employee Monitoring and Workplace Surveillance

Employee monitoring has long created legal, ethical, managerial, and technical challenges. Early Internet-era monitoring debates centered on productivity, acceptable use, email monitoring, web access, employer liability, and employee privacy expectations. Yerby (2013) argued that employee monitoring was already a complex legal and ethical issue because organizations monitor employees for productivity, security, legal liability, and acceptable use concerns, while employees often perceive monitoring as intrusive or unfair. The same work emphasized that monitoring programs require clear policies, employee education, and transparency regarding what is monitored and why.

Broader workplace surveillance literature confirms that monitoring is not only a technical practice. It is also an exercise of organizational power, risk management, and workplace control. Ball (2010) reviewed workplace surveillance research and described the proliferation of practices and controversies surrounding the monitoring of workers. Holland et al. (2015) found that electronic monitoring and surveillance can negatively affect trust in management, particularly in certain occupational contexts. Kayas (2023) synthesized workplace surveillance literature and identified organizational, managerial, and employee-level tensions surrounding surveillance practices. König (2025) further reviewed electronic monitoring at work as an established organizational psychology issue, showing that monitoring remains a current and active research concern rather than an outdated Internet-era debate. Ajunwa et al. (2017) argued that modern forms of worker surveillance raise serious legal and policy concerns because monitoring capabilities can expand faster than workplace privacy protections.

Security-focused monitoring also carries organizational risk (Yerby, 2013). Jeong and Zo (2021) examined opportunity-reducing techniques for preventing insider threats and found that hard controls can cause employees to feel intruded upon, reducing their perceived moral obligation to follow security policies and potentially increasing insider threat risk. This finding is important for agentic AI governance because organizations responding to attribution drift with broader surveillance may unintentionally weaken the trust and compliance environment they are trying to protect.

These findings are important because agentic AI may intensify existing monitoring tensions. Organizations may respond to agent-mediated risk by expanding surveillance of prompts, AI interactions, browser sessions, application activity, screenshots, keystrokes, file access, cloud actions, and endpoint telemetry. Such expansion may be understandable from a security perspective, but it also creates privacy, trust, and fairness concerns. The historical lesson from employee monitoring literature is that surveillance expansion is not a neutral or cost-free response to technological uncertainty. Monitoring may improve visibility while

simultaneously damaging trust, weakening morale, or producing unfair disciplinary decisions if the evidence does not distinguish human action from agent-mediated activity (Yerby, 2013).

Enterprise Monitoring, Insider Threat, and Credential-Based Attribution

Enterprise security monitoring frequently depends on linking observed activity to users, accounts, devices, systems, and contextual baselines. Unlike broader employee monitoring, which may focus on productivity, acceptable use, or workplace behavior, enterprise security monitoring is often used to support threat detection, incident response, insider threat investigations, compliance, and forensic reconstruction. Security information and event management platforms collect and correlate logs across systems. Endpoint detection and response and extended detection and response platforms observe endpoint activity, process execution, credential use, network connections, and suspicious behavior. User and entity behavior analytics tools attempt to detect anomalous behavior by comparing current activity against baseline patterns. Insider threat programs likewise depend on identifying actors, motives, opportunities, behaviors, and organizational context.

Insider threat literature illustrates the importance of actor identification and behavioral context. Homoliak et al. (2019) developed a structural taxonomy of insider threat incidents and emphasized the value of organizing incidents around information-gathering questions such as who, what, when, where, why, and how. Stanton et al. (2005) developed a taxonomy of end-user security behaviors, highlighting that organizational security depends not only on technical controls but also on user behavior and managerial ability to assess, influence, and constrain risky actions. CISA (2020) frames insider threat mitigation around defining, detecting, identifying, assessing, and managing threats.

Agentic AI complicates this environment because activity may no longer map neatly to human intent. A monitoring system may observe that a user account accessed a database, copied files, triggered scripts, modified source code, or sent messages. Historically, investigators would still consider malware, compromised credentials, automated scripts, and benign service activity. However, agentic AI adds another category: authorized or semi-authorized software acting on behalf of the user. The resulting activity may be neither malicious compromise nor direct human action. It may be delegated activity that appears in logs as human credential use.

This creates an evidentiary problem as much as a technical one. A log entry may accurately show that a credential, account, or process performed an action, but the record may not reveal whether the action was directly performed by a human, approved by a human, delegated to an agent, initiated by an agent through stored permissions, or produced by a background service. For insider threat and incident response teams, this distinction matters because the same observed action may imply different levels of culpability, negligence, compromise, or governance failure depending on the role of the agent.

Zero trust architecture also helps frame the issue. NIST SP 800-207 describes zero trust as a shift away from static network perimeters toward continuous evaluation of users, assets, and resources (Rose et al., 2020). Agentic AI introduces a complication for this model because the user, asset, and resource relationship may include an AI intermediary acting through existing credentials or delegated permissions. If zero trust requires continuous validation of identity and context, agentic AI requires organizations to ask whether the identity being validated represents a person, an agent, a system process, or some combination of these.

Agentic AI and Delegated Digital Work

Intelligent agents are not new. Wooldridge and Jennings (1995) described agents in terms of autonomy, social ability, reactivity, and proactivity. What is new is the integration of large language models, tool use, memory, planning, and orchestration into workplace systems that can operate across email, calendars, browsers, code repositories, databases, cloud platforms, and enterprise applications. Recent work on agentic AI describes systems capable of pursuing complex goals with reduced human oversight (Acharya et al.,

2025), shaping organizational and societal futures through autonomous capabilities (Hosseini & Seilani, 2025), and moving beyond narrow AI agents into systems with multi-agent coordination, persistent memory, and dynamic task decomposition (Sapkota et al., 2026).

Large language model research helps explain why these systems differ from traditional automation. ReAct demonstrated that language models can incorporate reasoning and actions, allowing systems to use external environments while updating plans (Yao et al., 2023). Generative agents research showed how agents can store memories, reflect, retrieve relevant information, and plan behavior in ways that simulate believable human-like activity (Park et al., 2023). Surveys of LLM agent planning describe task decomposition, plan selection, external module use, reflection, and memory as important components of autonomous agent behavior (Huang et al., 2024). Tool-learning agents further demonstrate the importance of integrating LLMs with external tools, APIs, and user-defined capabilities (Xu et al., 2025).

These features matter for enterprise monitoring because they expand the distance between user intent and system action. A user may provide a high-level instruction such as “summarize client risk and prepare a response,” while an AI agent retrieves documents, queries internal systems, drafts messages, creates files, and initiates workflows. The user may not observe every intermediate action. The agent may operate through the user’s authenticated session, delegated OAuth permissions, API tokens, browser automation, background services, or separate accounts. The resulting logs may capture activity, but not always the relationship between human intent, agent reasoning, and system execution.

Service Accounts, Daemons, and the Non-NoveltY of Automation

A credible governance analysis must avoid overstating novelty. Enterprise systems have long relied on daemons, services, scheduled tasks, management agents, service accounts, CI/CD pipelines, and orchestration tools. In these environments, a credential does not always equal a human user. These systems can create attribution and security problems, but their roles are often administratively defined, narrowly scoped, and documented around specific tasks. Agentic AI is more ambiguous because it may act as a general-purpose assistant, operate across multiple tools, adapt to user context, and perform tasks that resemble human knowledge work.

Agentic AI does not create the first non-human actor in enterprise systems. Instead, it expands the category of semi-autonomous actors and places them closer to human workflows, human credentials, and human accountability. The governance challenge is not simply automation. It is delegated, adaptive, and sometimes opaque automation operating in spaces where organizations traditionally expected human intent to be inferable from account activity.

Accountability, Responsibility, and Algorithmic Governance

The attribution problem also connects to longstanding accountability literature. Nissenbaum (1996) warned that computerized systems can erode accountability through barriers such as the problem of many hands, bugs, blaming the computer, and ownership without liability. Matthias (2004) described the responsibility gap that emerges when “learning automata” act in ways that are not fully predictable or controllable by human operators. Mittelstadt et al. (2016) argued that gaps between algorithmic design, operation, and ethical understanding can produce severe consequences for individuals and society.

These ideas are directly relevant to agentic AI in workplace monitoring. When an AI agent performs an action, responsibility may be distributed across the employee, vendor, employer, IT administrator, system designer, model provider, workflow owner, and policy environment. The employee may have initiated the task, but not the specific action. The employer may have approved the tool, but not the agent’s particular output. The vendor may have designed the system, but not the organizational context in which it operated. The monitoring system may have captured activity, but not intent.

Attribution drift is not simply a log-management issue. It is an accountability problem within socio-technical systems. Socio-technical systems theory is useful because the attribution problem does not arise

from AI agents alone (Baxter & Sommerville, 2011). It emerges from the interaction of human users, agentic tools, identity and access management systems, credentials, organizational policies, monitoring platforms, security analysts, vendors, and incident response procedures. A digital action may appear in a log as an account event, but its meaning depends on the surrounding socio-technical context: who authorized the agent, what permissions were granted, whether the user approved the action, how the agent executed the task, and whether monitoring systems preserved evidence of that relationship. Attribution drift reflects a system-level governance failure rather than a purely technical logging deficiency.

Attribution drift also extends accountability and responsibility-gap literature by focusing on the evidentiary conditions that precede responsibility assignment. Existing scholarship on accountability and responsibility gaps explains why it can be difficult to assign responsibility when automated or algorithmic systems contribute to harmful outcomes. Attribution drift narrows that concern to enterprise monitoring and cybersecurity governance. Before an organization can determine whether responsibility rests with an employee, employer, vendor, administrator, workflow owner, or model provider, it must first determine whether the observed activity was direct human conduct, agent-assisted conduct, delegated agent activity, system automation, or malicious activity. In this sense, attribution drift describes how responsibility gaps may be produced or intensified by weak identity separation, incomplete telemetry, unclear delegation, and insufficient forensic readiness.

This study does not attempt to resolve legal liability for agent-mediated harm. That question will likely depend on employment law, contract law, negligence standards, agency relationships, vendor obligations, regulatory duties, and future case law. Instead, the framework identifies a prior governance requirement: organizations need sufficient attribution confidence before they can make fair disciplinary, legal, security, or compliance judgments. Attribution drift helps explain why visibility into activity is not enough. Organizations must also preserve evidence about human intent, agent delegation, permission scope, approval events, tool use, and system context before responsibility can be evaluated.

Research Questions

This analysis is guided by three research questions for conceptual analysis:

RQ1: How do agentic AI systems complicate traditional assumptions of identity, intent, and accountability in enterprise monitoring?

RQ2: How does attribution drift affect employee monitoring, insider threat analysis, and organizational governance?

RQ3: What governance practices can organizations adopt to preserve attribution confidence while avoiding unnecessary surveillance expansion?

These questions were not tested through empirical data in this study. Instead, they guided the development of a conceptual governance framework, a taxonomy of human-agent activity, and propositions for future empirical research.

Methodology

This analysis used a conceptual framework methodology. Conceptual papers are appropriate when emerging phenomena require theoretical clarification, synthesis across literatures, and development of constructs that can guide later empirical work. The analysis was organized as an integrative conceptual

synthesis across five bodies of literature: employee monitoring and workplace surveillance; insider threat and enterprise security monitoring; agentic AI and LLM-based agents; algorithmic accountability and responsibility gaps; and AI governance and risk management.

Literature was selected based on its relevance to three analytic concerns: how digital activity is attributed to users or systems, how autonomous or delegated technologies complicate responsibility, and how monitoring practices affect governance, privacy, security, and forensic readiness. Peer-reviewed sources were used to ground the conceptual framework, while practitioner sources were used more narrowly to establish the organizational relevance and pace of agentic AI adoption. The paper does not claim that agentic AI has already produced a large body of documented legal or HR disputes. The literature remains sparse, particularly with respect to employee monitoring and attribution.

The framework development proceeded through three analytic steps. First, existing employee monitoring and insider threat assumptions were identified, especially the assumption that digital activity can be associated with users, credentials, devices, and intent. Second, recent agentic AI capabilities were examined to identify how planning, memory, tool use, delegated permissions, and background operation can disrupt those assumptions. Third, governance implications were organized into a taxonomy and propositions that can support future empirical testing.

Conceptual Framework: Attribution Drift

This paper introduces attribution drift as a conceptual construct for enterprise monitoring and AI governance. Attribution drift is the gradual erosion of confidence that observed digital activity can be reliably attributed to intentional human conduct when semi-autonomous AI agents operate under, alongside, or near human credentials.

Attribution confidence is the degree to which an organization can reliably associate observed digital actions with intentional human conduct, delegated agent activity, system-level automation, or malicious activity. High attribution confidence exists when logs, policies, identities, and technical controls make it reasonably clear who or what performed an action, under what authority, and with what level of human involvement. Low attribution confidence exists when digital activity is visible but the relationship between human intent, agent execution, and system behavior remains unclear.

The framework rests on a simple distinction: visibility is not the same as attribution. An organization may collect extensive endpoint, network, identity, and application telemetry while still lacking enough context to determine whether a suspicious action was directly human, agent-assisted, delegated, automated, or malicious. This distinction is central to the paper. The problem is not that organizations lack data or tools to monitor systems. The problem is that existing data may be poorly structured for reconstructing human-agent responsibility.

This distinction also clarifies the novelty of attribution drift. Existing accountability and responsibility-gap literature explains how responsibility becomes difficult to assign when outcomes are produced through complex technological systems. Attribution drift narrows that concern to the enterprise monitoring and cybersecurity context. The construct emphasizes the evidentiary degradation that occurs when human credentials, AI agent actions, delegated permissions, service accounts, and monitoring telemetry no longer provide a clear basis for interpreting observed behavior. Attribution drift therefore operates at the level of attribution confidence rather than final responsibility assignment. It describes the conditions under which organizations may have visibility into an event but insufficient confidence to classify the event as human-directed, agent-assisted, delegated, automated, or malicious.

Attribution drift can occur through several pathways. First, an agent may act under human credentials, causing logs to represent agent-mediated activity as user activity. Second, an agent may create or rely on background services, hidden accounts, or delegated tokens that are technically authorized but not clearly documented. Third, an agent may perform multi-step tasks that exceed the user's direct awareness. Fourth, a monitoring system may detect activity without understanding whether it reflects user intent, agent action, benign automation, or compromise. Fifth, investigations may reconstruct event sequences without sufficient evidence of the agent's prompts, reasoning, intermediate decisions, or tool calls.

The construct is intentionally distinct from full autonomy. If a system were fully autonomous and clearly identified as such, accountability might remain difficult, but the human-agent boundary would at least be visible. The harder problem is partial autonomy, where agent activity is embedded in ordinary user workflows and monitored through systems designed for human, system, or malicious activity categories. In that environment, organizations may over-attribute agent actions to employees, under-recognize agent-mediated risks, or respond by expanding surveillance in ways that harm trust and privacy.

Taxonomy of Human-Agent Activity

Table 1 presents a taxonomy of human-agent activity and attribution complexity. The taxonomy is intended to help organizations distinguish between different forms of AI involvement.

This taxonomy answers RQ1 by showing that agentic AI complicates identity, intent, and accountability differently depending on the mode of involvement. Human-directed AI use creates relatively limited attribution difficulty because the user remains close to the output and execution. Persistent and multi-agent operations create higher attribution complexity because actions may occur over time, across systems, and beyond the user's immediate observation.

Table 1. Taxonomy of Human-Agent Activity and Attribution Complexity

Mode	Description	Example	Monitoring Problem	Attribution Complexity
Human-directed AI use	A human directly prompts AI for a discrete output and decides how to use it.	Employee asks AI to draft a paragraph, summarize a policy, or suggest code.	Output may be AI-assisted, but system actions remain primarily human initiated.	Low
Agent-assisted action	AI recommends or prepares actions, but the human approves execution.	AI prepares an email reply, pull request, or workflow action that the employee reviews and submits.	Approval may be shallow, rushed, or treated as a rubber stamp.	Moderate
Delegated agentic activity	AI performs multi-step tasks across tools or systems on behalf of the user.	Agent retrieves files, queries systems, drafts responses, updates records, or creates tickets under a user session.	Logs may show user credentials while intent and execution are distributed across human and agent.	High
Persistent agent operations	AI agent or companion runs continuously through background services, delegated	Agent monitors inbox, manages calendar, updates documents, executes scripts, or performs recurring tasks.	Activity may be difficult to separate from user behavior, service behavior, or suspicious automation.	Very high

	tokens, or additional accounts.			
Multi-agent or orchestrated operations	Multiple agents coordinate tasks, share memory, or use different tools to accomplish organizational goals.	Agents coordinate research, reporting, scheduling, code changes, and system updates across applications.	Attribution may be distributed across agents, users, vendors, and orchestration layers.	Very high

Governance Propositions

The framework supports five propositions for empirical research and organizational analysis. These propositions are not tested hypotheses. Instead, they translate the conceptual framework into claims that can be examined through surveys, scenario experiments, case studies, technical audits, and policy analysis.

P1: As agentic AI systems gain persistence, tool access, and permission scope, attribution confidence decreases.

P2: When agentic AI activity is logged primarily through human credentials, organizations face increased risk of misattributing agent-mediated actions to employees.

P3: Attribution drift increases pressure for expanded employee monitoring unless organizations adopt agent-specific identity, logging, and governance controls.

P4: Transparent agent identity, permission separation, audit trails, and forensic readiness practices can preserve attribution confidence while reducing unnecessary surveillance expansion.

P5: Agent-specific monitoring can increase oversight of autonomous or delegated AI activity while reducing the need for intrusive monitoring of human employees.

Future research could operationalize attribution confidence through investigator confidence ratings, actor-classification accuracy, classification time, evidence requirements, and perceived fairness of disciplinary or security decisions. Agent autonomy could be measured through persistence, tool access, permission scope, human approval requirements, and background operation. Surveillance expansion could be measured through the number and intrusiveness of monitoring categories added after agent deployment, including prompt capture, screen recording, keystroke logging, browser telemetry, and AI interaction logging.

Together, the propositions frame attribution confidence as both an empirical construct and a governance design goal. If agents have distinct identities, scoped permissions, and auditable activity trails, organizations may monitor agentic systems more intensively while preserving more reasonable privacy expectations for human workers.

Discussion of Findings

The first major finding is that agentic AI complicates identity and intent in enterprise monitoring. Traditional monitoring environments already include non-human actors, but organizations often manage these through documented service accounts, device management policies, endpoint tools, and administrative controls. Agentic AI blurs these categories when agents operate through human sessions, browser automation, OAuth permissions, API tokens, or background services that are not easily understood by users or administrators.

The practical result is a weakening of the credential equals user assumption. In many systems, logs record that a user account performed an action. That record may remain technically accurate while being behaviorally incomplete. The user's credential may have authorized the action, but the user may not have directly performed the action, observed each step, or understood the agent's intermediate decisions. This distinction matters in HR investigations, insider threat analysis, legal disputes, compliance audits, and forensic reviews.

The issue is not that employees should be excused from all agent-mediated activity. Employees may still be responsible for selecting tools, granting permissions, approving actions, ignoring warnings, or failing to supervise delegated tasks. The problem is that responsibility may be shared, layered, or context-dependent. Organizations need evidence that distinguishes direct human action from agent-assisted or delegated action before making disciplinary, legal, or security determinations. In short, an account name in a log should be treated as the start of an investigation, not the end of one.

The second finding is that attribution drift affects insider threat analysis. Insider threat programs often distinguish malicious insiders, negligent insiders, compromised insiders, and accidental policy violators. Agentic AI introduces a possible additional category: agent-mediated activity that may appear negligent, suspicious, or malicious without being directly intended by the employee. For example, an agent might access more files than expected while preparing a report, upload sensitive data to an external service for summarization, execute an unsafe command, or generate a message that violates policy.

Monitoring tools may flag such actions correctly as risky. However, the interpretation of the alert becomes more complicated. An administrator or analyst must determine whether the action represents employee misconduct, compromised credentials, benign automation, misconfigured agent permissions, vendor behavior, inadequate policy, or normal but poorly documented agent functionality. Without agent-specific logs and identity markers, the organization may lack the evidence needed to make that distinction.

This creates a forensic readiness problem. Digital forensics depends on preserving evidence that can reconstruct events, actors, timelines, and intent where possible. Agentic AI may require new evidence categories, including prompts, agent plans, tool calls, intermediate observations, memory retrievals, approval events, delegated tokens, account creation events, and agent-specific logs. Without these artifacts, investigations may show what occurred but not whether the action was human-directed, agent-assisted, delegated, or autonomous within a permitted workflow. The result could be both over-attribution to employees and under-recognition of technical or governance failures.

The third finding is that attribution drift may push organizations toward broader employee surveillance. If organizations cannot distinguish human action from agent action, they may monitor more deeply. They may capture prompts, browser sessions, screenshots, keystrokes, AI chat histories, application windows, file contents, and communication metadata. Such measures may improve visibility, but they also increase privacy risk, employee stress, and trust concerns. Workplace surveillance literature shows that monitoring is not neutral. It can affect trust, job satisfaction, stress, and perceptions of fairness (Ball, 2010; Holland et al., 2015; Kayas, 2023; König, 2025; Siegel et al., 2022).

A governance-first approach should aim to avoid reflexive surveillance expansion. The objective should be attribution confidence, not total visibility. Organizations should collect the evidence necessary to distinguish human, agent, system, and malicious activity without defaulting to the most intrusive monitoring possible. Agent-specific identity, logging, and permission separation may allow organizations to monitor AI agents more closely while reducing the need to monitor every human interaction.

Legal and HR implications are important but should be treated as future research areas rather than the main focus of this paper. Agentic AI could create disputes involving employee discipline, wrongful termination, data leakage, discriminatory communications, unauthorized financial transactions, intellectual property

misuse, contract errors, cloud infrastructure, industrial control systems, healthcare systems, or safety-critical environments where agent-mediated actions produce harm.

The legal question in such cases may not be only whether the employee's account was involved, but whether the employee intentionally performed the action, negligently delegated authority, reasonably relied on an approved tool, or operated within an unclear governance environment. Did the employer approve the tool without adequate controls? Did the vendor create opaque background accounts or services? Did the monitoring system preserve sufficient evidence? Did policies clearly define acceptable use and responsibility for agent-mediated action? These questions illustrate why attribution drift cannot be resolved through account logs alone.

These questions are too broad for detailed legal analysis in this paper. Responsibility for agent-mediated harm will likely depend on the specific relationship among the employee, employer, vendor, administrator, workflow owner, and technical environment. The central point is not to assign liability in advance, but to show that organizations need evidence about prompts, approvals, permissions, configurations, service behavior, and security events before responsibility can be fairly evaluated.

Implications for Research and Practice

Governance Controls for Attribution Confidence

Table 2 maps key governance risks to recommended controls. The table is intended as a practical starting point for organizations adopting agentic AI in monitored environments and is consistent with AI risk management approaches that emphasize governance, measurement, management, and continuous oversight (National Institute of Standards and Technology, 2023).

Table 2. Governance Risks and Recommended Controls for Agentic AI Monitoring

Risk	Example	Governance Concern	Recommended Control
Agent uses human credentials	Agent accesses files, sends messages, or updates system via a user session.	Logs may over-attribute agent-mediated activity to the employee.	Use agent identity tagging, session markers, and tool-call logs.
Agent creates separate account or service	Installation creates a background account, daemon, service, or process.	Administrators may not know whether activity is authorized or suspicious.	Require transparent disclosure, inventorying, and approval of agent services and accounts.
Agent uses excessive permissions	Agent can access broad file stores, email, cloud tools, or repositories.	Over-permissioning increases data exposure and insider threat ambiguity.	Use least privilege, scoped tokens, time-bound permissions, and approval workflows.
Agent performs background actions	Agent works continuously or periodically without direct user observation.	Activity may be confused with malware, user behavior, or system automation.	Maintain agent-specific audit trails, recurring task registries, and notification requirements.
Monitoring flags employee for agent behavior	EDR, XDR, SIEM, or UEBA alert links behavior to user account.	HR or security response may misattribute intent.	Add investigative procedures to identify direct, assisted, delegated, and autonomous activity.

Organization expands surveillance broadly	Employer captures prompts, screenshots, keystrokes, and full sessions.	Privacy and trust may erode beyond what security requires.	Prioritize agent-specific logging, proportional human monitoring, data minimization, and governance review.
Agent-mediated incident requires investigation	Agent leaks data, executes unsafe code, or causes financial or operational harm.	Forensic evidence may lack prompts, plans, tool calls, approval records, and clear attribution.	Update forensic readiness plans to preserve agent logs, memory events, approval events, and permission changes.

Acceptable Use Policies and Employee Training

Acceptable use policies should be updated to address agentic AI specifically. Traditional policies often address email, Internet use, software installation, data handling, and acceptable use of company systems. Agentic AI policies should additionally define approved tools, prohibited data inputs, requirements for human review, permission boundaries, logging expectations, and responsibility for delegated actions. Employees should be trained to understand that approving an agent, granting permissions, or delegating a task may create accountability even if the employee does not manually perform each step.

Policy should also avoid unfairly shifting all risk to employees. If an organization encourages or requires use of AI agents, it must also provide clear guidance, approved tools, technical safeguards, and monitoring transparency. It is not reasonable to pressure employees to use agentic systems while leaving them uncertain about whether an agent-created account, background process, or delegated action will later be treated as suspicious or personally attributable.

Identity and Access Management

Identity and access management should distinguish human users from AI agents wherever feasible. Agent activity should not disappear into ordinary user sessions without markers. Organizations should consider separate agent identities, delegated credentials, scoped permissions, just-in-time authorization, session tagging, and agent-specific access reviews. These controls align with zero trust principles by improving context and continuous evaluation of who or what is requesting access (Rose et al., 2020). Where feasible, agent sessions should be marked through identity-provider claims, OAuth scopes, API token labels, device posture signals, or application-level metadata so downstream logs can distinguish human and agent-mediated activity.

The goal is not to eliminate agent access. The goal is to make agent access governable. If an agent reads files, queries a database, sends a message, or changes a configuration, the organization should be able to determine whether the action was human-directed, agent-assisted, delegated, or persistent background activity.

Monitoring and Forensic Readiness

Organizations adopting agentic AI should update monitoring and forensic readiness practices. EDR, XDR, SIEM, and UEBA tools may need new event categories for AI-mediated activity. Detection engineering should also account for agent-created services, unusual delegated-token use, tool-call chains, background task execution, and agent actions that resemble normal user behavior. Incident response playbooks should ask whether an AI agent was involved, what permissions it had, what prompts or goals initiated the action, whether a human approved execution, what tool calls occurred, and whether the action was expected by policy.

Forensic readiness should include retention of agent logs, tool-call records, permission grants, prompts where appropriate, approval events, system account creation, and configuration changes. Privacy and data minimization remain important. Organizations need enough evidence to reconstruct high-risk events, but not every prompt or interaction should be recorded indefinitely. The balance should be attribution confidence through agent-specific evidence collection, not unlimited surveillance of human workers.

Taken together, the taxonomy, propositions, and governance recommendations address the three research questions. They show that agentic AI weakens traditional assumptions about identity, intent, and accountability when agents operate through human sessions, delegated permissions, background services, additional accounts, API tokens, or tool integrations. They also show that attribution drift complicates employee monitoring, insider threat analysis, HR discipline, legal accountability, compliance review, and forensic reconstruction. The recommended response is not broader surveillance by default, but transparent agent identity, permission separation, scoped credentials, agent-specific audit trails, updated acceptable use policies, AI-aware incident response, and forensic readiness practices that preserve attribution confidence while limiting unnecessary human surveillance.

Limitations

This paper is conceptual and does not present empirical data. The proposed taxonomy and propositions should be tested through surveys, interviews, scenario experiments, case studies, and technical audits. The paper relies partly on practitioner reports to establish the pace of agentic AI adoption because academic literature on enterprise agent deployment remains relatively new. These sources are useful for context but should not be treated as substitutes for empirical organizational research.

The legal analysis is intentionally limited. Agentic AI may raise questions involving employment law, contract law, negligence, agency law, intellectual property, discrimination, data protection, and sector-specific regulation. Those issues deserve focused legal scholarship. This analysis instead concentrates on cybersecurity governance, attribution, employee monitoring, and surveillance implications.

The proposed framework also assumes enterprise environments with some level of centralized monitoring, identity management, and security governance. Smaller organizations, bring-your-own-device environments, unmanaged endpoints, and informal AI use may face different attribution challenges. Future research should examine how attribution drift appears across organizational size, sector, maturity, and regulatory environment, especially in lower-maturity environments where identity management, endpoint governance, and centralized logging may be less developed.

Future Research

Several avenues for future research follow from this framework. First, attribution perception research could examine how employees, administrators, managers, auditors, and investigators assign responsibility for agent-mediated actions. Survey studies and scenario-based experiments could test whether responsibility is attributed differently across direct human action, agent-assisted action, delegated agentic activity, persistent agent operations, and multi-agent workflows. These studies should also examine whether judgments of blame, negligence, trust, and acceptable monitoring change across legal, ethical, financial, reputational, operational, and safety-related harm scenarios.

Second, technical attribution research could examine whether current EDR, XDR, SIEM, identity, and cloud monitoring tools can reliably distinguish human actions from agent actions. Future technical research should also examine whether agentic systems can intentionally or unintentionally obscure monitoring signals, evade attribution, or generate telemetry that resembles legitimate human activity. Such studies could identify gaps in logging, session tagging, token management, tool-call recording, agent memory retention, approval capture, and forensic evidence preservation. This research could also test whether agent-

specific identity and audit trails increase attribution confidence without requiring broader surveillance of human workers.

Third, policy and governance research could analyze how acceptable use policies, AI governance frameworks, insider threat programs, and incident response procedures are being updated to address agent-mediated activity. Longitudinal studies could examine whether organizations adopt agent-specific controls or instead respond to attribution drift through broader employee surveillance. Future governance research should also examine emergency disablement mechanisms for high-risk agent behavior, including technical kill switches, workflow suspension procedures, and escalation protocols. Sector-specific research should examine high-risk environments such as healthcare, finance, education, defense contracting, industrial control systems, and cloud infrastructure, where attribution drift may carry heightened legal, financial, or safety consequences.

Conclusion

Agentic AI creates a new governance challenge for enterprise monitoring because it blurs the boundary between human intent, automated execution, and delegated digital action. The issue is not that automation is new. Enterprise systems have long relied on daemons, service accounts, scheduled tasks, background agents, and privileged monitoring tools. The issue is that agentic AI introduces semi-autonomous systems that may act as extensions of human users while operating through credentials, services, permissions, and workflows that monitoring systems may not clearly distinguish.

This study introduced attribution drift as the erosion of confidence that observed digital activity can be reliably attributed to intentional human conduct when semi-autonomous agents operate under, alongside, or near human credentials. The proposed taxonomy distinguishes human-directed AI use, agent-assisted action, delegated agentic activity, persistent agent operations, and multi-agent operations. The analysis also developed propositions and governance recommendations intended to preserve attribution confidence without defaulting to excessive surveillance.

Full autonomy is not required for this problem to matter. Partial autonomy may be more difficult to govern because responsibility remains shared, intermittent, and difficult to reconstruct after an incident. Organizations adopting agentic AI should update monitoring, identity, acceptable use, incident response, and forensic readiness practices. More precise, agent-specific monitoring may help preserve security without expanding surveillance of human workers. Otherwise, organizations may find themselves collecting more data while understanding less about who or what actually did the work.

Acknowledgements

AI tools supported brainstorming and literature exploration. The author wrote, revised, and verified the manuscript and remains responsible for its originality, accuracy, and integrity.

References

- Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals: A comprehensive survey. *IEEE Access*, 13, 18912-18936. <https://doi.org/10.1109/ACCESS.2025.3532853>
- Ajunwa, I., Crawford, K., & Schultz, J. (2017). Limitless worker surveillance. *California Law Review*, 105(3), 735-776.
- Amazon Web Services. (n.d.). Model feature attribution drift violations. Amazon SageMaker Developer Guide. <https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-monitor-model-attribution-drift-violations.html>
- Ball, K. (2010). Workplace surveillance: An overview. *Labor History*, 51(1), 87-106. <https://doi.org/10.1080/00236561003654776>
- Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9), Article 404. <https://doi.org/10.3390/fi17090404>
- Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1), 4–17. <https://doi.org/10.1016/j.intcom.2010.07.003>
- Cybersecurity and Infrastructure Security Agency. (2020). Insider threat mitigation guide. U.S. Department of Homeland Security. <https://www.cisa.gov/resources-tools/resources/insider-threat-mitigation-guide>
- Holland, P. J., Cooper, B., & Hecker, R. (2015). Electronic monitoring and surveillance in the workplace: The effects on trust in management, and the moderating role of occupational type. *Personnel Review*, 44(1), 161-175. <https://doi.org/10.1108/PR-11-2013-0211>
- Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight into insiders and IT: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. *ACM Computing Surveys*, 52(2), Article 30. <https://doi.org/10.1145/3303771>
- Hosseini, S., & Seilani, H. (2025). The role of agentic AI in shaping a smart future: A systematic review. *Array*, 26, Article 100399. <https://doi.org/10.1016/j.array.2025.100399>
- Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., & Chen, E. (2024). Understanding the planning of LLM agents: A survey. *arXiv*. <https://arxiv.org/abs/2402.02716>
- Jeong, M., & Zo, H. (2021). Preventing insider threats to enhance organizational security: The role of opportunity-reducing techniques. *Telematics and Informatics*, 63, Article 101670. <https://doi.org/10.1016/j.tele.2021.101670>
- Kayas, O. G. (2023). Workplace surveillance: A systematic review, integrative framework, and research agenda. *Journal of Business Research*, 168, Article 114212. <https://doi.org/10.1016/j.jbusres.2023.114212>

- Kent, K., & Souppaya, M. (2006). *Guide to computer security log management* (NIST Special Publication 800-92). National Institute of Standards and Technology.
<https://doi.org/10.6028/NIST.SP.800-92>
- König, C. J. (2025). Electronic monitoring at work. *Annual Review of Organizational Psychology and Organizational Behavior*, 12, 321-342. <https://doi.org/10.1146/annurev-orgpsych-110622-060758>
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175-183. <https://doi.org/10.1007/s10676-004-3422-1>
- Microsoft. (2025). The 2025 annual work trend index: The frontier firm is born. Microsoft WorkLab.
<https://www.microsoft.com/en-us/worklab/work-trend-index/2025-the-year-the-frontier-firm-is-born>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21. <https://doi.org/10.1177/2053951716679679>
- National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2, 25-42. <https://doi.org/10.1007/BF02639315>
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, Article 2. <https://doi.org/10.1145/3586183.3606763>
- Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology.
<https://doi.org/10.6028/NIST.SP.800-207>
- Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion*, 126, Article 103599.
<https://doi.org/10.1016/j.inffus.2025.103599>
- Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, I., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI.
<https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>
- Siegel, R., König, C. J., & Lazar, V. (2022). The impact of electronic monitoring on employees' job satisfaction, stress, performance, and counterproductive work behavior: A meta-analysis. *Computers in Human Behavior Reports*, 8, Article 100227. <https://doi.org/10.1016/j.chbr.2022.100227>
- Stanton, J. M., Stam, K. R., Mastrangelo, P., & Jolton, J. (2005). Analysis of end user security behaviors. *Computers & Security*, 24(2), 124-133. <https://doi.org/10.1016/j.cose.2004.07.001>
- Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2), 115-152. <https://doi.org/10.1017/S0269888900008122>

Xu, W., Huang, C., Gao, S., & Shang, S. (2025). LLM-based agents for tool learning: A survey. *Data Science and Engineering*, 10, 533-563. <https://doi.org/10.1007/s41019-025-00296-9>

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of the 11th International Conference on Learning Representations.

Yerby, J. (2013). Legal and ethical issues of employee monitoring. *Online Journal of Applied Knowledge Management (OJAKM)*, 1(2), 44-55.
https://www.iiakm.org/ojakm/articles/2013/OJAKM_Volume1_2pp44-55.php