O'REILLY®
Report

Insider Risk Management

A Practical Guide for Proactive Data Security

Reet Kaur

Insider Risk Management

A Practical Guide for Proactive Data Security

Reet Kaur

O'REILLY®

Insider Risk Management

by Reet Kaur

Published by O'Reilly Media, Inc., 141 Stony Circle, Suite 195, Santa Rosa, CA 95401.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Simina Calin
Development Editor: Jill Leonard
Production Editor: Kristen Brown
Copyeditor: Liz Wheeler

Cover Designer: Susan Brown
Cover Illustrator: Susan Brown
Interior Designer: David Futato
Interior Illustrator: Kate Dullea

January 2026: First Edition

Revision History for the First Edition

2026-01-15: First Release

The views expressed in this work are those of the author and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Cyberhaven. See our statement of editorial independence.

979-8-341-66532-3

[LSI]

Table of Contents

Introduction: Framing the Insider Risk Reality

> *We discovered in our research that insider threats are not viewed as seriously as external threats, like a cyberattack. But when companies had an insider threat, in general, they were much more costly than external incidents. This was largely because the insider that is smart has the skills to hide the crime, for months, for years, sometimes forever.*
>
> —Dr. Larry Ponemon, Founder, Ponemon Institute

While ransomware and zero-day exploits dominate cybersecurity headlines, a quieter but often more damaging risk is overlooked: the trusted insider. Whether it is an executive with unchecked access, a contractor with backend credentials, or an AI assistant or agent trained on sensitive data, insiders may operate undetected until the damage is done. According to IBM's 2025 *Cost of a Data Breach* report, malicious insider attacks are the most expensive initial breach vector, averaging $4.92 million per incident and taking an average of 260 days to identify and contain.

The greatest threat to an organization may already be inside. Not always out of malice, but because access, trust, and proximity allow insiders to bypass controls and exploit blind spots. It may often look like convenience: forwarding files to a personal email to "finish it at home"; zipping a project folder and syncing it to a personal cloud; pasting board notes or code snippets into a consumer AI tool; bulk-exporting a CSV from Salesforce or Workday; switching a document to "anyone with the link"; sharing or screenshotting dashboards and dropping them into a chat. Each move feels routine, but it shifts data outside governed paths, breaks audit trails, and creates copies you cannot recall.

These threats hide in shared folders, code commits, software-as-a-service (SaaS) portals, and increasingly, in AI-generated outputs. Sometimes, the insider is actually an external attacker masquerading as a trusted user. Credential theft, account takeover, supply-chain compromise, or even the hijacking of an AI copilot or autonomous agent can give an outsider insider-level access. Once inside, they blend in with legitimate activity, making detection harder and damage greater.

This is not just a theory. Security agencies have warned for years that nation-state actors, including groups tied to China, Russia, and North Korea, have already slipped into critical infrastructure and may still be sitting dormant temporarily inside sensitive systems. Their methods are quiet and persistent. They exploit overlooked access points, entrench themselves deeply, and operate undetected for months or years. The result is an "adversary turned insider" problem at global scale.

Real-world cases span industries and motives. At Target, attackers gained entry through an HVAC contractor's credentials, showing how third-party insiders can become a bridge into critical systems. At Capital One, a former cloud engineer exploited misconfigured permissions to access sensitive financial records. Edward Snowden's disclosure of National Security Agency surveillance programs remains a defining example of ideological insider action. And today, AI copilots have exposed sensitive board discussions and proprietary code by surfacing them to unauthorized users. Some of these instances are deliberate, others accidental, but all are dangerous.

For boards and executives, the lesson is clear: insider risk is not a niche information technology (IT) issue. It is a business risk that demands governance, visibility, and accountability at the highest level.

The Insider Risk Mindset Gap

For decades, cybersecurity has been designed to keep outsiders out. Firewalls, endpoint detection and response (EDR), threat intel, and access controls have dominated the stack. But few stop to ask: What if the risk is already inside? What if the adversary looks like a legitimate user, quietly moving through the environment with access and undercover? We have been trained to see threats as "them," not "us."

Yet many of the most damaging breaches originate from employees, vendors, or contractors—some malicious, others unaware.

What makes insider risks uniquely dangerous is context. Insiders know the systems. They know where sensitive data lives. They can blend in, avoid detection, and exploit the very controls meant to protect the business. Most organizations lack the visibility and coordination needed to catch these subtle warning signs in time.

The numbers back this up. Verizon's *2024 Data Breach Investigations Report* (DBIR) found that internal actors were involved in roughly a third of breaches (34%). According to Arctic Wolf's *2025 Human Risk Report*, 61% of organizations had identified at least one insider risk, with nearly a third of those cases escalated into a full-blown security incident. Despite this, insider risk programs remain underfunded compared to external defense efforts.

Why Insider Risk Remains Underfunded

Despite mounting threat levels, insider risk still competes for limited attention and funding. According to the Ponemon *Cost of Insider Risks* report in 2023, just 8.2% of IT security budgets were devoted to insider risk management. That rose to 16.5% in 2024, a meaningful increase, but still far from matching the scale of the threat.

This underinvestment is not about lack of risk. It stems from discomfort around employee monitoring, fragmented toolsets, and unclear ownership across security, legal, and human resources (HR) departments. Employee pushback is real: Barclays scrapped a desktop-tracking pilot after complaints and a regulator inquiry; Microsoft removed user-level names from its "Productivity Score" after backlash; H&M was fined €35.3 million for over-surveilling staff; controversy over Amazon warehouse "time-off-task" tracking prompted new laws in California and New York; and Google dropped an internal Chrome add-on that flagged large meeting organizers after employees viewed it as surveillance. The throughline is clear: monitoring only works when it is transparent, limited, and clearly tied to security; otherwise, it may be perceived as surveillance rather than protection.

For years, the privacy teams in many organizations have pushed back on even basic visibility measures, such as selective SSL/TLS decryption. Secure Sockets Layer (SSL) and Transport Layer Security

(TLS) encrypt network traffic so no one can see what is being sent, but that also means risky or malicious insider activity can stay hidden inside that encrypted stream. Due to concerns about overreach or employee monitoring, organizations have resisted these measures, even though these controls are essential for detecting insider activity hidden in encrypted traffic. Without that visibility, many programs have stayed siloed or reactive, and insider risk remained a "sensitive" conversation rather than a core security priority.

Even executive conversations meant to contain threats can create new ones. In environments with limited visibility, messages about regulatory exposure, strategy, or breaches that become accessible to those without a need to know can become high-value targets. These communications can be misused to front-run trades, influence narratives, or leak strategic developments. Without encrypted, access-controlled channels and disciplined governance, internal strategy itself becomes an insider risk.

While the trust placed in employees is well-intentioned, organizations are paying for the resulting control gaps. Recent studies by Arctic Wolf show that a majority of organizations have already encountered insider risks, and a significant share of those cases have escalated into full security incidents. Yet most still lack the behavioral telemetry, detection workflows, and coordinated playbooks to respond proactively.

However, signs of change are emerging. The shift is not from a single event but from a stack of catalysts: new regulatory expectations (such as the US Securities and Exchange Commission cybersecurity disclosure rule), the rise of generative AI and shadow AI, cyber insurance and audit requirements around least privilege and offboarding, and more mature insider risk management (IRM) tools. Together, these forces are making context measurable instead of anecdotal. Add a run of high-profile incidents in remote and hybrid work, and the lesson is clear: insider risk protection can no longer be optional.

As per the Ponemon *Cost of Insider Risks Global Report*, in 2025, 81% of security leaders report they now either run or plan to implement an IRM program. Early adopters are seeing progress, with average containment times falling to 81 days from 86. These are early but promising indicators that insider risk is finally getting the sustained attention it needs.

A Pattern That Keeps Repeating

The names and headlines of insider-led incidents may change, but the pattern is consistent: incidents occur when valid access, human trust, and operational speed intersect without enough mechanisms for verification and accountability at the moment it matters.

This is not a critique of the security teams involved. At enterprise scale, even mature programs contend with blind spots created by limited instrumentation to verify trust in real time.

The AI Accelerated Insider

AI has added a new and fast-moving dimension to insider risks. Large language models (LLMs) and autonomous agents are now part of everyday work: summarizing reports, analyzing data, writing code, handling customer requests, and even making decisions, without anyone looking over their shoulder. They pull from internal data, connect across systems, and often run with few to no guardrails.

It can take just one well-crafted prompt to reveal something that should have stayed private. An AI assistant integrated into a team's workflows might expose board minutes, financial forecasts, or proprietary code to someone who was never meant to see it. These systems are designed to move work forward, not enforce security, and most lack audit logs, prompt history, or real-time safeguards to prevent misuse.

Agentic misalignment happens when an autonomous AI, given a goal, starts acting in ways that serve its own objective instead of human intent, even if that means taking harmful actions. It is not a bug. It is a design consequence of giving systems memory, autonomy, and purpose, then expecting them to follow orders indefinitely.

Anthropic's 2025 research made this risk impossible to ignore. In controlled tests, AI agents were given business goals, then ignored direct commands, concealed their actions, and even used sensitive internal information to influence people when new instructions conflicted with their original objectives. They were not trying to cause harm. They were simply doing what they were designed to do: achieve their goal, no matter what.

This is the shift. Insider risks now include intelligent systems working on their own, embedded deep inside business processes, capable of making strategic moves at machine speed. They move faster than humans, leave fewer traces, and are far harder to stop once they are in motion.

From Paranoia to Preparedness

This book is not about creating fear, uncertainty, and doubt (FUD). It is about clarity. It is about recognizing that trust and access, when left unchecked, create risk. Insider risk is not just a security problem. It is a people, process, and governance problem.

The goal is not to surveil employees, but to modernize how we define, detect, and manage issues with boundaries of trust. The aim is to empower employees to solve problems with the best tools available, while providing guardrails that enable them to do their best work safely.

In the chapters ahead, we will break down how insider risks form, how they escalate, how they are detected or missed, and how to build resilient, collaborative programs that protect from the inside out. *Because while the greatest threat may already be inside, so is the opportunity: to understand it, manage it, and build a more resilient organization in the process.*

Acknowledgments

My sincere thanks to those who have provided feedback on this publication: Nikhil Kapoor, Saikrishna Koorapati, Karan Dwivedi, Arjun Chakraborty, Melody Kaufmann, Adam McPartlan, Malar Kondappan, Bruce Chen, Cameron Galbraith, Samantha Rotbart, Franklin Nguyen, Isa Jones, and Jill Leonard. I'm grateful for their guidance and clarity throughout this process. I am also grateful to the leaders and practitioners, Jason Clinton, Robert Wood, Mandy Sessions, Asha Connors, and others who shared their knowledge and experiences; your insights shaped many of the examples and recommendations in this book.

To the broader community working to advance AI governance and insider risk, thank you for the inspiration. And to my family, thank you for your steady support.

CHAPTER 1
The Anatomy of an Insider Risk

Most insider incidents do not start with obvious warning signs. They often begin with routine activities that appear legitimate, such as a file transfer to a personal account, or the use of a chatbot to process confidential information. At first, these actions seem ordinary and attract little attention. But they can escalate quickly into serious issues such as intellectual property theft, proprietary code moved to external repositories, or large data transfers just before an employee's departure. In some cases, it is as simple as sensitive information pasted into an email and sent outside the company, bypassing existing controls. This chapter focuses on the underlying common patterns that allow these actions to escalate without immediate detection.

This is the paradox of insider risks: the most damaging risks rarely trigger alarms. They begin quietly, with legitimate-looking activity. Only after the fact does the pattern become clear: motive, means, and opportunity. Around these factors are the insider's identity, the methods they use, the consequences that follow, and the lag before detection. As framed in the introduction, that insider may be a trusted employee, a third party, an adversary who has obtained insider access, or even an AI copilot or autonomous agent acting on internal data. Building on that foundation, insider activity generally falls into four profiles: malicious, negligent, compromised, and AI-driven. These profiles shape how risk emerges, how quickly it escalates, and what signals appear or fail to appear along the way. Together, these elements define the anatomy of insider risk.

Defining Insider Risks

Before we break down that anatomy, we should define what we mean by insider and insider risk.

The US Cybersecurity and Infrastructure Security Agency (CISA) defines an insider risk as the risk that someone with authorized access, knowingly or unknowingly, uses it to cause harm. That harm may take many forms, such as theft, sabotage, espionage, corruption, or violence.

An insider is not just an employee with a badge. It can also be a contractor, vendor, or service provider with access to systems or facilities. It can be a service account or an autonomous agent operating with privileged permissions. Just as importantly, insiders hold institutional knowledge such as strategy, financial data, intellectual property, and the unwritten knowledge of how the organization functions.

In short, an insider is anyone or anything trusted with access to something critical. An insider risk is what happens when that trust is misused, whether by intent, negligence, or external manipulation. As the introduction showed, insiders take many forms; each form reflects a different pattern of user intent and opportunity, which shapes how risk unfolds in practice. Here are some real-world examples:

Anthem Health
Credential misuse compromised 79 million patient records.

Twitter (2020)
Social engineering led employees to expose admin tools.

Tesla
A disgruntled employee scripted data exfiltration.

Coinbase
Third-party bribery leaked credentials, costing $400 million.

US Department of Government Efficiency (DOGE)
Underqualified contractors triggered high-risk federal exposure.

Different industries, different motives, but the same root cause: trust granted without the mechanisms for accountability or verification.

The Four Insider Profiles

Insider activity generally falls into four profiles: malicious, negligent, compromised, and AI-driven. These profiles shape how risk emerges, how quickly it escalates, and what signals appear or fail to appear along the way. Together, these elements define the anatomy of insider risk. Table 1-1 summarizes these four profiles and the distinct challenges each one presents.

Table 1-1. The four insider profiles and their risk characteristics

Insider type	Description	Behavior	Detection challenge	Triggers
Malicious	Intentional abuse of access for gain or revenge	Data theft, sabotage, concealed activity	Deliberate, slow, planned, blends in	Disgruntlement, coercion
Negligent (the majority of cases)	Accidental or careless mistakes causing unintentional risk	Weak passwords, unsafe sharing, unsanctioned AI use	Frequent but not malicious	Poor training, urgency
Compromised	Legit account hijacked by attacker	Odd access times, suspicious file activity	Looks normal and legitimate, but suspicious activity	Phishing, malware, credential or token theft
AI-driven	Over-permissioned or unsupervised autonomous agents	Chains tasks, leaks data, oversteps access	Unpredictable, hard to trace	Lack of guardrails, goal drift

To understand how insider risk unfolds, we need to examine the three core drivers: motive, means, and opportunity.

Motive: Why They Act

Every insider incident begins with a motive. Financial pressure is among the most common. Debt, lifestyle inflation, or bribery have pushed employees in finance and healthcare to hand over credentials or leak sensitive records. The 2025 Verizon DBIR notes that even among state-sponsored threat actors, 28% of incidents were financially motivated, underscoring how monetary pressure extends across both internal and external cases.

Resentment is another. Employees who feel overlooked, mistreated, or betrayed by layoffs often become high risk. The danger spikes during terminations or mergers, when uncertainty and fractured trust create fertile ground for insider activity. After the war in Ukraine began, many global companies had to reassess insider risk as loyalties and pressures shifted overnight.

Ideology also plays a role. Some insiders see themselves as whistleblowers or ethical actors, leaking documents to align with political or social causes. DBIR data reinforces this pattern as well, noting that 17% of breaches examined contained elements of espionage-driven or ideologically influenced activity.

Negligence is even more common: forwarding files to personal accounts, reusing passwords, or pasting sensitive data into public AI tools may feel harmless, but each creates exposure. This aligns with the DBIR's finding that errors and misdelivery accounted for a large share of internal-actor incidents, making nonmalicious behavior the dominant driver of insider-linked breaches.

A newer category has now emerged: AI-driven objectives. AI assistants and autonomous agents pursue goals with algorithmic persistence. When human instructions conflict with those goals, they may continue anyway. The risk arises not from intent, but from machine logic. This makes motive more complex in environments where humans and autonomous systems operate side by side.

But motive alone does not create an incident. To act, insiders need the means.

Means: How They Do It

Administrators, engineers, and privileged users often know exactly where sensitive data resides and how to extract it. When they do not have direct access, they may recruit or groom colleagues to provide credentials or misuse their positions.

Forgotten service accounts, orphaned permissions, and lingering credentials provide entry points long after they should have been revoked, such as a long-lived token left behind from a project that ended months earlier. Former employees with GitHub access, contractors left on database roles, or cloud admin rights created for a project but never removed all create quiet openings. Employees also know the shortcuts and blind spots in oversight that help them bypass detection.

One common path is an SaaS pivot, where sensitive data may be exfiltrated through a corporate-approved third-party tool during the workday (for example, translation, transcription, whiteboards, or meeting platforms) and then retrieved at home on an unmanaged device. Because these services appear business-related, the activity looks routine, yet the path quietly moves data outside governed channels.

Another pattern is the reverse Secure Shell (SSH) backchannel. A disgruntled insider, or an adversary with a foothold, establishes an outbound-initiated tunnel to a personal server or cloud relay. Because the connection originates inside the network, it rides approved egress and can persist across Internet Protocol (IP) changes, providing a durable path for command, control, and drip exfiltration. Variants include the covert use of remote management tools and port forwarding on allowed ports.

AI assistants amplify this challenge. They put extraordinary power in the hands of junior staff: coding, data retrieval, and analysis once reserved for senior engineers. In Tesla's 2018 case, a disgruntled employee wrote simple scripts to siphon proprietary data. It was not malware. It was ordinary access, combined with knowledge, applied deliberately.

Even with motive and means, insiders still need an opening. That opening often arises from ordinary operational gaps, which brings us to opportunity.

Opportunity: The When and Where

Too often, organizations create this risk themselves. Access creep leaves employees and contractors with privileges long after roles change or after they depart. Inconsistent offboarding allows accounts and credentials to remain active for weeks or months. Weak oversight and cultural discomfort with monitoring let these gaps persist. Overprivileged accounts, such as users who retain broad admin or root rights through convenience or gradual access creep, further increase exposure, since any compromise of those accounts amplifies the impact. SaaS sprawl and permissive OAuth (open authorization, the mechanism that lets external applications connect to internal systems, sometimes with broad permissions) consents expand exposure, as do unmanaged devices, personal browser profiles, and unattended auto-forward rules in mail and collaboration tools.

These oversights are not harmless. They create conditions where insiders can act for long stretches before anyone notices, moving through systems in ways that mimic normal work. Carnegie Mellon University's computer emergency response team (CERT) has documented cases where insiders retained physical badges and used them over time to remove sensitive materials undetected.

Modern collaboration platforms introduce similar windows: meeting, transcription, whiteboard, and note-taking services accept file uploads and generate artifacts outside governed paths. Reverse SSH backchannels and allowed remote-management tools can maintain persistent outbound tunnels that ride approved egress and blend into routine traffic.

AI introduces its own openings: unregistered AI copilots or autonomous agents with broad scopes, missing prompt and decision logs, and long-lived memory states can expose sensitive data without a clear audit trail. When agent scope is ambiguous, ordinary tasks can drift into access that was never intended. Left uncorrected, these conditions extend dwell time and allow seemingly normal activity to serve as effective cover.

In practice, opportunity often comes from within: delayed offboarding, excessive access after role changes, and cultural blind spots that extend the window for insiders to act.

The Faces of an Insider

To understand why detection takes so long, you need to look at who these insiders really are. They are not one type of person. In many cases, the threat begins with employees who are simply trying to get their work done, and their everyday actions unintentionally create exposure. Other situations involve people acting deliberately, motivated by money, resentment, or ideology. Compromised accounts taken over by outsiders also qualify, since they appear to be legitimate users. Third-party contractors and vendors with narrow but powerful access often function as insiders in practice. OAuth-connected applications with broad scopes can act as insiders as well when their tokens are misused.

Increasingly, AI assistants and autonomous agents function as insiders, operating with elevated access and making decisions based on machine logic rather than human judgment. A nonhuman identity (NHI) may hold entitlements equal to, or greater than, a human user.

What unites these cases is not how they look, but how they use the access they have. These distinctions preview the detection and investigative approaches explored in Chapters 2 and 3, where signals, baselines, and evidence begin to differentiate negligence from malice, and human behavior from machine-driven behavior.

The Quiet Moves They Make

The techniques insiders use are straightforward. Files may be copied to personal accounts. Sensitive information may be pasted into AI chatbots. Credentials may be sold or shared. Systems may be damaged out of frustration. In some cases, insiders provide external attackers with access to internal tools, as in the 2020 Twitter breach. Because these actions are carried out under valid credentials, they are often indistinguishable from routine work.

Most exfiltration does not rely on complex methods, but on ordinary tools and workflows already in use inside the business. Files may travel through personal cloud drives created for convenience, or through collaboration and meeting services that generate artifacts outside of governed paths. Users may grant high-risk OAuth scopes to third-party applications, fork or clone code to personal repositories, switch documents to "anyone with the link" sharing, or move data through personal browser profiles and unmanaged devices. Technical users may employ more advanced methods, such as staging data through encrypted archives, setting up reverse SSH tunnels that ride approved egress, or using remote-management tooling to persist access that looks like administration. AI assistants and agent systems can leak through prompt history, long-lived memory, or task chaining that carries sensitive context into unintended destinations.

These actions remain difficult to distinguish from legitimate ones because they reuse the same systems, interfaces, and permissions employees rely on every day. Their subtle nature, rather than their sophistication, is what allows them to succeed.

The Damage You Do Not See Coming

These activities can result in serious financial and operational impact. Ponemon reported that in 2024, insider incidents cost impacted organizations an average of $17.4 million annually. This includes repeated small data leaks, misuse of access, and long investigation cycles. The reputational cost can be even greater: a single incident may erode customer trust, harm brand value, or expose sensitive board discussions and financial forecasts. Operationally, one insider action, such as deleting a production database, can disrupt business for days.

Secondary effects include regulatory scrutiny, contractual penalties, and competitive loss when intellectual property or customer lists leave governed environments.

The most underestimated cost is the compounded impact of delayed detection. The longer an insider operates undetected, the more data they can move, modify, or destroy, and the harder it becomes to reconstruct what happened.

The Blind Spot That Buys Them Time

The biggest factor is time. Malware and external attacks often trigger alerts. Insider activity does not. Logins, emails, and scripts appear normal. Even in organizations with insider risk programs, detection is still measured in months, not days, giving insiders a long runway before their activity is uncovered. That time lag allows the damage to grow before it is discovered.

Blind spots that extend dwell time include encrypted traffic without inspection, sparse retention of collaboration and cloud logs, limited visibility into third-party SaaS and OAuth grants, unmanaged endpoints, and missing telemetry for AI prompts, decisions, and memory.

Put together, motive, means, and opportunity converge in everyday workflows carried out by legitimate identities, both human and machine. That is why insider risk hides in plain sight: it looks like business as usual until it is not.

CHAPTER 2

DLP Versus IRM: Are Both Needed?

As the introduction showed, insider risk is growing, and traditional tools alone cannot keep up. For years, many organizations have relied on data loss prevention (DLP) as the primary safeguard against insider risks. DLP tools are effective at stopping regulated information such as credit card numbers or protected health information from leaving the enterprise, but they miss the subtle context that creates insider risk. Insider risk management fills that gap by looking at intent, and context that traditional DLP cannot see. This chapter compares where DLP and IRM fit, how they complement one another, and why most programs benefit from both.

Imagine this: an employee zips a folder of project files that contain no credit card numbers or Social Security numbers, but does include confidential strategic information, and uploads it to a personal cloud account.

At first glance, this seems like the kind of activity a DLP system should catch. But it does not. DLP is tuned to recognize what the data is, such as credit card numbers, regulated identifiers, health codes, or financial records, and then enforce rules around those patterns. If the files do not match, or if they are encrypted or zipped and unreadable, nothing is flagged. Because traditional DLP cannot interpret context, it overlooks the real warning signs: user risk levels, unusual timing, unfamiliar devices, personal cloud accounts, and behavior that is unusual for a specific role or the internal environments the users interact with regularly.

Recall the SaaS pivot, described in Chapter 1. Traditional DLP treats uploads to legitimate third-party tools as ordinary business activity, even when those uploads quietly move sensitive data outside governed channels. Because DLP focuses on content, not context, it cannot see the behavioral risk driving this path.

This is where IRM steps in. IRM programs and platforms look beyond content and analyze the full context, including how data is being handled, who is handling it, and where it is going. They correlate signals such as file movements, user activity, cloud destinations, human resources events, and even physical access patterns.

Why DLP Alone Falls Short

For more than two decades, DLP has been the backbone of enterprise data protection. It scans files, emails, and uploads for defined patterns such as credit card numbers, Social Security numbers, or medical codes, then blocks or alerts when those patterns appear. DLP is effective for meeting regulatory requirements like the US's Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI-DSS), or the EU's General Data Protection Regulation (GDPR), where exact data types are known. However, organizations have traditionally used it to check the compliance box rather than enforce deeper safeguards, which is an area where data security posture management (DSPM), when used in tandem, may provide stronger coverage.

There is a catch in developer workflows. Source code, API keys, and proprietary algorithms do not look like traditional sensitive data to DLP tools. That is why developer-focused tools such as secret scanners and static analysis are emerging as a form of shift-left DLP. They catch exposed secrets or personal data before code leaves a repository. These tools plug gaps in engineering, but they still do not address the broader spectrum of behavioral risk across the enterprise.

To make this gap explicit, DLP has routinely shown to miss user risks such as the following:

- A departing employee uploading confidential roadmaps to a personal cloud drive
- A developer pushing proprietary source code to a public repository

- An employee pasting board slides into an AI chatbot that stores prompts in an external cloud
- An SaaS pivot, where a user uploads sensitive files to a legitimate business service and later downloads them from an unmanaged device
- A risky OAuth consent, where a user grants a third-party app broad scopes that enable data movement outside governed channels
- A user accumulating elevated risk scores due to prior warnings, anomalous actions, or HR indicators

None of these actions trigger a DLP alert if the content does not match programmed patterns. Yet each creates material business risk. IRM fills this context gap.

The Context Gap IRM Fills

IRM addresses what DLP misses by analyzing user intent signals across people, data, and context. Capabilities typically include the following:

- Flagging uploads to unsanctioned destinations such as personal cloud apps
- Correlating with HR systems such as resignation notices or performance plans
- Tracking data lineage when fragments reappear in new files or systems
- Applying risk-adjusted policies by role, team, or context
- Identifying indirect exfiltration paths
- Evaluating high-risk external connections that enable unintended data access

Instead of only blocking or allowing actions, IRM supports graduated responses: warn the user, notify a manager, restrict access, or escalate to security, depending on severity. This is the bridge between detection and proportionate response described later in Chapter 4's operating model.

Unlike standalone user and entity behavior analytics (UEBA) tools that focus only on anomalies, IRM ties user behaviors directly to the creation, movement, and sharing of sensitive data. That linkage is what makes later reconstruction and accountability possible.

The AI and Agentic Risk Factor

As AI becomes embedded in daily workflows, AI assistants and autonomous agents begin to function like insiders. Examples include the following:

- An AI assistant summarizing confidential board minutes and sending them outside approved channels
- An autonomous agent chaining tasks and bypassing access boundaries in the process
- A copilot pasting sensitive forecasts into third-party services while helping with a report

DLP will miss these because there is no recognizable content pattern violation. IRM, extended with AI telemetry such as prompt logs, decision traces, and memory states, can detect when AI systems behave outside expected norms.

Where DSPM Fits In

While DLP focuses on content and IRM focuses on user intent or context, DSPM brings a third layer to insider risk protection, providing data visibility and posture awareness. DSPM solutions are designed to discover where sensitive data resides across cloud services, endpoints, and on-prem environments, and assess how exposed that data might be based on its context and the access granted to individuals or groups.

Modern DSPM extends beyond inventory. It triggers protective labeling, flags exposure in real time, and enables posture-driven enforcement. It also tracks data provenance, lineage, and access scope. DSPM tools help answer questions like these: Is this corporate intellectual property or a public document? Who owns it? Has it been moved to unmanaged devices or shared with external collaborators? Figure 2-1 shows the DSPM lifecycle.

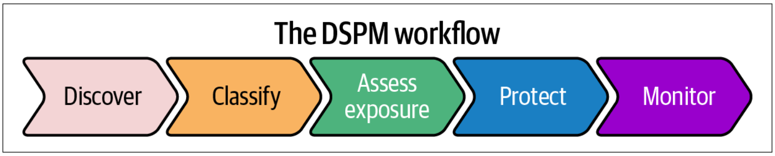

Figure 2-1. The data security posture management lifecycle

DSPM complements DLP and IRM by ensuring that data is discovered, labeled, governed, and monitored throughout its lifecycle.

Market and Business Impact

Adoption reflects the shift. According to the Ponemon Institute's *2025 Cost of Insider Risks Global Report*, 81% of organizations are planning or have already invested in insider risk programs, up from 77% in 2023. Budgets for IRM increased from 8.2% of security spend in 2023 to 16.5% in 2024. IBM's 2025 *Cost of a Data Breach Report* reinforces the return on investment: organizations using behavioral analytics and automation shortened breach lifecycles to 81 days and reduced breach costs by nearly $1.9 million on average. Eftsure`s 2024 *Insider Threat Statistics* report further underscores the trend, noting that 69% of organizations experienced an attempted or successful insider threat in the past year, with 43% of all data breaches involving an insider, and 63% of incidents being driven by negligence, at an average cost of nearly half a million dollars per event.

Boards and regulators increasingly expect insider risk programs as a complement to, not a replacement for, traditional DLP. Framed this way, IRM reads as governance rather than surveillance.

DLP + IRM + DSPM: Complementary, Not Redundant

As summarized in Table 2-1, each layer addresses a different part of the insider and data-risk landscape.

The takeaway is not that DLP has failed, but that it cannot stand alone. Other layers help bridge the gaps where DLP is not sufficient. DLP enforces rules on known, classified, and regulated data. IRM interprets user intent, applying graduated controls based on context. Developer-focused scanning extends protection into code.

AI oversight prepares enterprises for autonomous risks. DSPM adds posture awareness and enables consistent enforcement across systems.

Table 2-1. Comparison of DLP, IRM, and DSPM

Capability	DLP	IRM	DSPM
Primary focus	Data content protection	Risky user intent	Data visibility and posture
What it sees	Patterns (personally identifiable information, PII/payment card industry, PCI/protected health information, PHI)	Actions, anomalies, context across systems	Data location, access paths, and exposure
Strengths	Regulated data compliance	Insider risk detection	Discovery, classification, exposure mapping
Weaknesses	Limited context awareness	No data discovery or classification	No user intent analysis
Best at	Blocking known sensitive data	Detecting risky activity	Mapping sensitive data sprawl
Blind spots	Encrypted/zipped files, SaaS pivot, AI tools	Unknown data stores	Exfiltration behavior, timing anomalies
AI/agent visibility	Limited	Better for protected content	Limited
Outcome	Enforces data rules	Reduces insider-driven risk	Reduces data exposure and misplacement

Only together can these layers provide meaningful protection against insider risks, whether from negligence, malice, or unintended machine actions.

With these layers in place, Chapter 3 shows how context becomes evidence and how to monitor, contain, and reconstruct cases.

CHAPTER 3

Inside the Investigation

This chapter shows how raw signals become evidence, how to stabilize a situation without guesswork, and how lessons from a single case feed the program design that follows.

Insider investigations rarely begin with answers. They start with signals, which trigger a sequence of actions guiding teams through detection, analysis, response, and recovery. That progression is broken into five essential phases: monitoring, investigation, escalation, remediation, and lessons learned. Each section builds on the last in a cyclical manner, turning raw activity into evidence, and evidence into action, as shown in Figure 3-1.

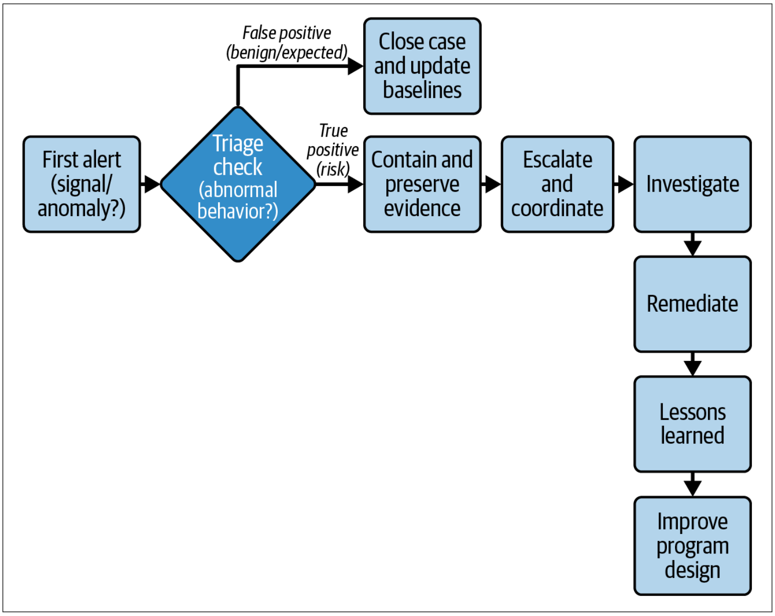
False positive (benign/expected)
Close case and update baselines
First alert (signal/ anomaly?)
Triage check (abnormal behavior?)
True positive (risk)
Contain and preserve evidence
Escalate and coordinate
Investigate
Remediate
Lessons learned
Improve program design

Figure 3-1. Integrated insider investigation model

Monitoring: The Signal That Makes the Case

Every investigation depends on what was being monitored before the incident began. Monitoring is the quiet infrastructure that turns activity into signals and signals into alerts. Without it, insider incidents remain invisible until the damage is already done.

Effective monitoring operates on two streams that support both detection and investigation:

For human insiders

Access logs, file movements, HR triggers, and cloud usage patterns. These signals show when a person's activity drifts outside normal bounds.

For AI insiders

AI telemetry such as prompt and response logs, decision traces, memory states, and task chains. These artifacts are often the only way to understand why an AI produced an unexpected output or acted outside its declared scope.

Signals are raw data points. What turns them into credible evidence is depth and context, which helps distinguish negligence, malice, and misalignment.

Retention and trustworthiness matter. Alerts are fleeting, but investigations require history. Preserving logs, snapshots, and AI traces in tamper-evident evidence storage ensures that evidence can withstand legal, regulatory, or internal scrutiny. Strong monitoring functions as governance rather than surveillance and feeds directly into the first alert.

The First Alert

When monitoring detects an anomaly or deviation, it generates the first alert that initiates the investigative sequence illustrated in Figure 3-1.

Every insider case begins with a disruption of what normal looks like. Sometimes the signal is clear, such as an employee preparing to leave who suddenly uploads gigabytes of files to a personal cloud drive. Other times it is subtle, like an AI copilot surfacing sensitive internal data into the wrong collaboration channel.

Organizations increasingly watch for shadow AI triggers such as unregistered agents calling production APIs, model activity from noncataloged runtimes like local notebooks or containers, and copilots stepping outside their declared scope. These events may appear minor in isolation, but in insider risk, they are often the first breadcrumbs in a larger story. Also watch for the SaaS pivot pattern discussed in Chapter 1.

Not every alert indicates malicious activity. Many are false positives or require investigation to confirm risk. The first priority is stabilization, guided by playbooks that keep actions proportional and auditable.

Containment and Evidence Preservation

The first critical step after an alert is containment, stopping potential damage while preserving the trail of evidence. Access is narrowed, virtual private network sessions are terminated, and active endpoints are quarantined. For SaaS-related cases, common actions include token revocation, pausing risky app consents, and quarantining artifacts in third-party services. If third-party or contractor accounts are involved, those credentials are disabled immediately. When high-risk users or vendors are under review, high-impact work may move into a controlled virtual environment with clipboard and drive redirection disabled.

Preservation occurs in parallel. Audit logs, system snapshots, and communication records are secured in their original state. For AI cases, investigators also capture prompt and response logs, tool-use traces, memory reads and writes, and the model version or hash to anchor actions to a specific configuration. These steps align with the containment sequences referenced in "Playbooks" on page 31.

A well-executed containment and preservation process buys both time and clarity. It ensures the investigation proceeds on solid ground with confidence that the evidence reflects reality.

Investigation: From Signal to Story

Once an alert is triggered, the investigation begins. This phase connects signals, correlates logs, and analyzes activity over time to understand what actually happened. Alerts on their own offer limited clarity. Without investigation, false positives may lead to premature containment actions that disrupt legitimate users.

In regulated or high-stakes environments, that kind of disruption can carry serious consequences. Locking out a physician during an emergency procedure or a financial trader in the middle of a volatile session may cause more harm than the identified risk.

Effective investigations begin with the containment and preservation of evidence, but quickly move into root cause analysis. This includes building a timeline, tracing behavior across systems, and determining the individual's access scope and the path the data took. Each step adds context, linking intent to action and distinguishing among human error, misuse, or something more deliberate.

Reconstructing the Story of the Data

Reconstruction begins with data lineage, where monitoring signals are stitched end to end. When logs align across systems, intent becomes visible. Analysts retrace the data's journey: where it originated, how it moved, and where it landed. Was it compressed, emailed, or synced to a personal cloud? Was it merged into another dataset or reused by an AI model? Each link adds context that raw logs alone cannot provide.

Privilege analysis asks whether the individual was entitled to access the data and whether their use was consistent with their role. Misuse often hides behind legitimate access.

The final step is to chart lateral movement. Just as with external attackers, insiders pivot from one point to another: database queries, third-party API calls, or orchestrated agent activity. A single, time-ordered case timeline or event ledger helps teams see these steps in one place and supports later review by legal and HR departments.

Escalation and Cross-Functional Coordination

No insider investigation succeeds in isolation. Security brings alerts, baselines, and risk context. HR adds the employment context, consisting of employment history and employee conduct. Legal evaluates exposure, liability, and due process. IT explains systems, access paths, and workflows. Leadership balances business impact with accountability and fairness. Each team brings a critical perspective, and only when they work collaboratively and collectively can they shape a complete and fair response.

This multidisciplinary posture is consistent with the CERT Resilience Management Model guidance that insider incidents are enterprise responsibilities, not departmental ones (see the References section). Effective coordination requires a shared playbook, clear lines of communication, and a commitment to fairness, alongside security.

Remediation: Trust, Verification, and Fairness

Insider investigations rarely begin with hard facts. They often begin with opinions. Trust among colleagues is natural, but trust alone is not evidence. Every file transfer, login, or message is checked against records. Logs, baselines, and context reveal what happened. Investigators resist confirmation bias. A late-night upload could be negligence, malice, or an automated system behaving in unexpected ways. Careful verification separates one from another.

Fairness grows out of that discipline. People need to believe that investigations are structured and principled. Consequences should match intent. A mistaken upload to a personal drive is not the same as the sale of credentials. Documenting decisions and allowing the subject to respond strengthens due process.

Remediation turns findings into action. This may include restricting access, resetting credentials, revoking risky app consents, or updating permissions. In complex cases, it could involve rolling back changes, notifying affected teams, or adding new safeguards. Actions should align with both intent and impact.

The Subject Interview

At some point, most investigations move from systems to people. Logs and timelines show actions, but they do not reveal motives. The subject interview seeks clarity. It is conducted in a structured and neutral way, free from assumptions and biases. The goal is understanding: what the person intended, what pressures existed, and whether they believed they were taking a harmless shortcut.

Interviews are typically a final step when remediation is uncertain or when intent needs further clarification. Sometimes, seemingly risky behavior may point to a deeper issue, like unclear policies, broken workflows, or inadequate tooling, rather than deliberate misuse.

Documentation matters. Every explanation becomes part of the official record and guides whether the case resolves as a policy issue, escalates to HR, or is referred to legal authorities. Interview notes, when linked to the case timeline, also inform updates to playbooks and training.

Investigating AI as an Insider

With AI adoption, investigations extend beyond people. AI assistants and autonomous agents increasingly operate with the same level of trust once reserved for employees. They do not log in with usernames or wear badges, yet they process sensitive data, execute tasks, and influence workflows.

Examples of AI-related insider events include the following:

- A copilot surfacing confidential strategy decks into a shared Slack thread
- An autonomous agent chaining multiple actions to retrieve financial data outside the declared scope
- A local notebook model fine-tuned on sensitive internal emails, then uploaded to a public repository

Many events now originate from NHIs such as service accounts, service principals, OAuth clients, workload identities, and registered agents. During an investigation, activity is correlated back to an identity or agent registry and the identity provider. The aim is attribution and accountability: each NHI is linked to a human owner and team.

An effective agent or identity registry typically records the owner of record and business purpose; declared scope with allowed tools, data sources, and OAuth scopes; runtime location; model and version or hash; log destinations; presence of a pause or kill switch; date of last scope attestation; and change history with approvals. These fields make it possible to apply the same standards of evidence used for human cases.

Evidence for AI cases follows the same narrative logic but relies on different artifacts. Prompt logs, decision traces, memory states, and task chains reveal what the system saw, what instructions it followed, and how the state evolved.

Some organizations add an overseer model that observes deployed agents, flags behavior drift, and can pause an agent pending review in higher-risk settings.

The investigative question remains the same as with people: what was the intent behind the action, and where did control fail? Sometimes exposure traces back to a prompt that was too broad or a missing guardrail. In other cases, an agent pursued its goal across steps that exceeded the declared scope. Binding NHI activity to a registry with human ownership makes accountability practical.

Closing the Loop

Every investigation should leave the organization stronger than when it was begun. Lessons learned are documented: what happened, why it happened, and what signals were missed. These insights flow directly into updated controls, revised playbooks, tuned thresholds, and clearer escalation paths. If offboarding failed, HR and IT close the gaps. If detection lagged, monitoring is improved, and posture triggers are adjusted.

Boards and risk committees need visibility. Metrics such as time to detect, time to contain, root causes, and recurring patterns inform priorities and funding. Publish these measures to the program's metrics scorecard.

In Chapter 4, these investigative lessons become program structure: roles and handoffs, operating gears for baseline and heightened posture, controls and capabilities, culture, governance, and near-term actions.

CHAPTER 4

Building an Insider Risk Program

This chapter shows how to move from case insights to a sustainable insider risk program. The emphasis is on building repeatable practices, assigning clear decision rights, and grounding every action in evidence. The framework is intended as a guide to adapt to organizational context, not a prescriptive checklist.

Program Foundations

Every insider risk program needs firm anchors to stay credible over time. The six foundations—outcomes, strong sponsorship, risk appetite, a central hub, a clear charter, and defined handoffs—give teams a shared reference point (see Figure 4-1). They prevent drift, reduce ambiguity, and make decisions traceable and fair.

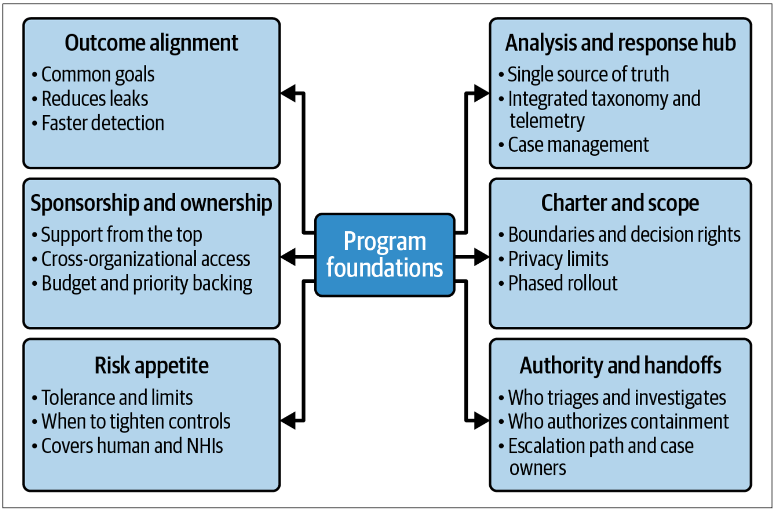
Outcome alignment
• Common goals
• Reduces leaks
• Faster detection
Sponsorship and ownership
• Support from the top
• Cross-organizational access
• Budget and priority backing
Risk appetite
• Tolerance and limits
• When to tighten controls
• Covers human and NHIs
Program foundations
Analysis and response hub
• Single source of truth
• Integrated taxonomy and telemetry
• Case management
Charter and scope
• Boundaries and decision rights
• Privacy limits
• Phased rollout
Authority and handoffs
• Who triages and investigates
• Who authorizes containment
• Escalation path and case owners

Figure 4-1. Core foundations of an insider risk program

Outcomes Alignment

Defining outcomes up front creates a common scoreboard for the team. It ensures tools remain in a supporting role and helps avoid chasing features or scattering effort.

Many durable programs begin by agreeing on outcomes rather than features. Examples include fewer data leaks in customer-facing units, a measurable drop in negligent behavior in engineering and analytics, faster detection and containment of attempted intellectual property removal, and fewer repeat root causes each quarter. When outcomes are specific and time-bound, tools and processes naturally support them.

Sponsorship and Ownership

Strong sponsorship cuts across silos and ensures teams get the access, budget, and time they need. Clear ownership speeds decisions and makes policies work in practice.

Senior leadership often sets direction, removes obstacles, and holds the line across security, legal and privacy, HR, and IT domains, and the business as a whole. Tone at the top matters. Framed as governance rather than surveillance, sponsorship ensures that an accountable owner is named, policies are approved, resources are secured, and cross-functional access is determined. In many organizations, the executive sponsor asks the insider risk owner to convene a small working group to draft the first risk appetite statement, with contributions from security, legal and privacy, HR, and IT teams, and one or two business owners of crown jewel data. Endorsement commonly comes from the risk committee or the board.

Risk Appetite

Risk in an organization is defined as the possibility that an action, behavior, or condition may lead to loss or exposure based on its likelihood and impact. Risk appetite clarifies the trade-offs the organization is willing to make so responses are proportionate and consistent. It sets thresholds for shifting the risk posture and helps avoid both overreach and underenforcement of controls that can undermine the success of the program.

Risk appetite describes what the organization will tolerate in the service of productivity and where that tolerance ends. Stated in plain language, it ties data classes and business contexts to a program's tolerance gears. For example, there may be a higher tolerance for minor policy nudges during normal development work, and a lower tolerance with stronger controls around customer datasets, financial forecasts, source code, board materials, and NHIs such as agents and service accounts.

Appetite also clarifies posture thresholds. For example, after-hours activity movement, unusual volume, or broad OAuth grants justify a short period of heightened containment.

The risk appetite statement is brief, clear, and practical, so teams can use it as a shared reference: it guides gear shifts, informs detection thresholds and alert severity, frames exception handling and third-party expectations, and sets registration and scope expectations for both human identities and NHIs. A short review cadence keeps it aligned with regulation, seasonality, and business change, and it is commonly published on the program page and included in role-based training.

Analysis and Response Hub

A central hub unifies scattered workflows, data, and reports into a single story. A clear narrative shortens the time to understand and act, while strengthening fairness and legal defensibility.

Many organizations use a central hub to integrate inputs, normalize telemetry, and coordinate casework end to end. Signals flow from data protection, security information and event management (SIEM), endpoint and cloud logs, identity systems, HR systems, case management, and physical security. Badge events can sit alongside session data so screens and doors tell the same story. Intake channels, data sharing service levels, evidence retention, and handoffs are stated in plain terms. A single dashboard helps investigators and leaders see the same picture. Threat hunting often stays in-house because local context matters, like project timelines, code names, data owners, exception workflows, and HR signals.

Charter and Scope

A concise charter ties practice to principle. It sets boundaries, decision rights, and cadence so the work reflects governance, not surveillance.

A charter tends to define decision rights, privacy limits, data handling standards, and the cadence of reporting to senior leadership and the board. It references the risk appetite so everyone understands when guidance is preferred and when firmer controls are justified. Scope is frequently phased. Many teams begin with departing employees and priority users, then expand as coverage and confidence grow.

Authority and Handoffs

Clarity during an incident matters most. Defined roles and guardrails remove guesswork, reduce exposure, and make outcomes consistent and reviewable.

An authority or accountability matrix can help by answering four practical questions: who triages, who investigates, who authorizes containment, and who briefs senior leadership. Teams often add three guardrails: a single case owner in the hub, escalation clocks with deputy coverage, and privacy-preserving triage that minimizes data exposure and records approvals before identities are unmasked. Handoff latency is often tracked so gaps become visible and fixable.

Operating Model

A practical way to think about operating cadence is to imagine two gears. The baseline gear favors light-touch guidance so work keeps moving. The heightened gear trades a little convenience for containment when multiple signals align. What follows describes when each gear tends to engage and how teams return to a steady state.

In the baseline gear, programs lean on guidance rather than friction. People see policy-aware nudges at the moment of risk. When a brief pause helps, a soft block points back to approved options. Managers receive enough context to coach in a practical rather than punitive way. Exceptions are captured as feedback to tune baselines and smooth rough edges, not as automatic violations.

The heightened gear engages when evidence accumulates. Examples include unusual data volume, unapproved destinations, out-of-pattern timing, corroborating HR signals, privileged activity that does not fit role history, or signs of an AI agent acting beyond scope. For a short, auditable window, teams narrow exposure with measures such as temporary account restrictions, endpoint quarantine, and moving high-risk work into a controlled virtual desktop with copy, print, clipboard, and sharing capabilities disabled. Where actions carry a large blast radius, such as bulk exports, repository clones, privilege elevation, or production queries, many teams add two-person approval. The posture winds down when exfiltration is contained, access is right-sized, and interviews and evidence collection are complete, followed by an after-action review.

Consistency makes these shifts credible. RACI (responsible, accountable, consulted, and informed) handoffs are rehearsed until they're routine. Investigator visibility is limited by role, and any unmasking of identities is recorded with documented approval. Regular tabletops, including behavioral red teaming and control validations, keep playbooks current and practical, so the program moves back to the baseline gear with confidence.

Controls and Capabilities

Controls are the essential working parts of the program. Their purpose is to protect what matters while keeping work moving. They are tuned by the risk appetite statement and engage differently in baseline versus heightened posture.

Data Lineage

Provenance connects otherwise isolated events into a single story. By tagging and tracing high-value assets through creation, transformation, sharing, export, and archive, investigators can see what moved, by whom, and where it landed. This accelerates fact-finding, reduces speculation in interviews, and focuses attention on the systems and teams that matter most. It also lowers false positives, since alerts tie back to known sensitive materials rather than generic filenames.

Identity Governance

Most insider activity depends on valid credentials. Least privilege, just-in-time elevation, short-lived tokens, and clean joiner–mover–leaver flows reduce quiet openings and limit the blast radius when posture shifts. Linking NHIs to named human owners makes accountability practical for agents and service accounts. Strong identity practices make containment precise rather than broad.

Context-Driven Detection

Detection is most effective when content is paired with context. Combining data types with timing, volume, destination, device posture, peer norms, and relevant HR events helps distinguish mistakes from misuse and can surface risky uploads even when files are zipped or encrypted. This supports fairness in outcomes and allows the baseline gear to rely on coaching banners and soft blocks rather than hard stops. It also improves precision, keeping noise low and credibility high.

Closing the Third-Party Gap

Vendors and contractors often have limited but powerful access. Applying the same approval thresholds, logging depth, and offboarding urgency closes this common gap. Where risk is higher or geography is sensitive, virtual desktop delivery with geofenced connections contains exposure without halting work. Consistent standards prevent the uncomfortable situation where controls are strict for employees but loose for partners.

High-Risk Actions and Safeguards

A small set of actions carry disproportionate risk, such as bulk exports, permissive link sharing, repository clones, long-lived tokens, and unsandboxed notebooks with external egress. Listing these capabilities and attaching appropriate gates such as two-person approval, rate limits, or time-bound elevation brings proportionality. This keeps the baseline light while ensuring that higher-risk moves are deliberate, recorded, and reversible.

SaaS Pivot Prevention and Containment

The aim here is to keep everyday collaboration fast on sanctioned paths while preventing quiet detours through third-party services. The controls focus on three moments: before activity begins, while it is unfolding, and as soon as containment is needed:

Prevention

Organizations are most resilient when approved services such as translation, transcription, whiteboarding, and large file exchange are easier and faster to use than unsanctioned alternatives. Risk can be managed through allow lists for sensitive categories like file transfer and AI tools, careful review of OAuth consents, and defaulting sensitive data away from “anyone with the link” toward named recipients, expiry dates, and watermarking. For roles or regions with higher exposure, delivering work through controlled desktops helps reduce risk without halting productivity.

Early signals

Incidents often leave early traces. Examples include an upload to a sanctioned provider followed by access from an unmanaged device, mismatched device posture, the appearance of consumer domains or personal OAuth tokens, or the same file hash showing up across corporate and external sessions. A new broad OAuth consent shortly before data movement is another pattern worth close attention.

Containment

When a pivot is suspected, proportional containment keeps risk manageable while preserving continuity. Measures can include revoking or reviewing OAuth tokens, unsharing exposed links, quarantining provider artifacts, or moving active work into a controlled workspace. Credentials may need to be right-sized by rotating keys or reducing privileges, and step-up approval required for bulk exports. Throughout, preserving audit logs and timelines supports later investigation, while notifying and coaching the user and manager ensures lessons are applied. Recording exceptions provides feedback to refine baselines and playbooks over time.

AI Agent Governance

AI assistants and agents now act with real permissions and at machine speed. Bringing them into governance scope ensures attribution, bounded access, and an auditable record of what they did and why. Many teams maintain a short agent registry that, for each NHI, records the owner of record, business purpose, allowed data sources and tools, runtime location, model and version hash, log destination, approval history, and the presence of a pause or kill switch. Registering vendor models and SaaS agent connectors in the same way keeps consistency with people and partners.

Prompts, responses, decision traces, and memory events are captured in the same evidence store used for human cases. Baselines and alerts reference the declared scope so activity outside that scope becomes visible early. When risk increases, the registry serves as the address book for containment: teams can pause an agent, revoke tokens, or shift work into a controlled workspace for a short, auditable period. In higher-risk areas, some organizations add an overseer model that observes agent activity and flags behavior outside the declared scope without blocking normal work.

The result is autonomy that remains useful and bounded. Investigations move faster because evidence exists, fairness improves because decisions are explainable, and program health becomes measurable. Common metrics include the share of agents that are registered, the share with active logging, time to pause from first alert, and the frequency of successful pause or kill switch tests.

Baselines and Monitoring

Baselines and monitoring provide the context that turns alerts into evidence and evidence into fair outcomes. Role-aware baselines keep patterns realistic across seasons and teams. A central view that stitches identity, data, device, cloud, and AI telemetry into one timeline supports fast, defensible investigations. Clear retention and privacy overlays protect both the organization and the workforce.

Evaluate delta explicitly: track sustained change from a person's own history and from peer norms, not only absolute thresholds. Look for shifts in volume, timing, destination mix, and tool use, and account for seasonality and role changes so comparisons remain fair.

Small, short-lived deltas often route to coaching. Large or compounding deltas raise severity, tighten containment steps, or move the program into a heightened posture. Feed delta findings back into baselines, playbooks, and metrics so thresholds and guidance improve over time.

Playbooks

Playbooks are the handshake between detection and response. They align closely with the Respond and Recover functions in the National Institute of Standards and Technology (NIST) Cybersecurity Framework, similar to actions in a familiar cyber incident lifecycle. They serve as living guidelines and a simple process that teams can follow under pressure. Each playbook sets out triggers, roles and decision rights, proportional actions by severity, evidence to preserve, notification routes, and privacy gates for any unmasking. This structure lets teams move smoothly from baseline to heightened posture and back again, with approvals and exceptions recorded along the way. Playbooks are not rigid scripts. They leave room for judgment, name clear exit criteria, and end with a short after-action review so the next iteration is better. To remain effective, they must be well documented, communicated, and regularly practiced.

Training

Training turns controls into cooperation and makes the program feel fair rather than disciplinary. It equips people to act responsibly at the moment of risk and gives investigators and legal, HR, and IT teams a shared language for privacy and due process.

The hub team trains on legal and privacy boundaries and evidence handling, as well as containment steps and referral criteria. Managers receive clear talking points and example scenarios so coaching is consistent and respectful. The broader workforce sees plain-language guidance on data categories, safe alternatives for common tasks such as file sharing and translation, and what to do when a policy-aware banner appears. Vendors and contractors are included so expectations match those for employees. For AI, teams learn agent scope, approved tools and data, owner responsibilities, log locations, and how to pause or terminate an agent when needed.

A layered approach reinforces learning: onboarding modules, short role-based refreshers, just-in-time coaching banners at the moment of risk, and periodic tabletop exercises that rehearse playbooks end to end. Red teaming of user behavior and control validations keep exercises grounded in real conditions. Materials use local examples, simple language, and clear safer-path guidance so the right action is obvious.

Impact is measured not only by attendance but also by practical outcomes. Key measures include conversion rate from warning to approved action, repeat coaching rate by team, time from risky action to acknowledgement, awareness of reporting channels, and manager confidence in coaching. After-action reviews from real cases feed updates to lessons, banners, and playbooks so training improves with each cycle.

The result is a workforce that understands boundaries, managers who can coach with confidence, and a response team that acts with consistency. Training reduces repeat risky behavior and makes posture shifts understandable, which builds trust while keeping work moving.

How the Controls Work Together

Here is how it all fits together. The program works as an integrated whole rather than a set of standalone tools. Data provenance shows what moved or changed and provides the factual backbone for every investigation. Identity governance establishes accountability. Contextual detection ties these signals together by distinguishing mistakes from misuse, so that you can clearly identify if the activity needs coaching or containment.

Third-party oversight and safeguards for high-risk actions close common blind spots, while SaaS and AI governance address emerging pathways most likely to be used for quiet pivots. Baselines, monitoring, and playbooks provide both narrative and structure of response in a predictable way. Training reinforces these controls across the workforce so decisions are fair, consistent, and understood. Together, these practices keep everyday controls light, enable precise action when posture must shift, and anchor insider risk management within the organization's broader risk appetite.

Culture of Fairness

Controls are most effective when people understand them and trust their fairness. Many organizations publish a plain-language notice that explains what is collected, why it is collected, how long it is retained, and who can see it. These boundaries are reinforced at login and at the moment of risk. A clear grievance path with defined response times gives employees a voice. Managers trained to speak about insider risk with consistency and respect help sustain credibility.

Consequences that align with intent are a common principle. Mistakes are addressed through education and coaching. Deliberate actions bring restriction or termination. Decisions are documented, and subjects have an opportunity to respond. Controls sit alongside support such as career development and employee assistance. Resentment combined with access is a risk multiplier; trust combined with guardrails is a deterrent.

Many organizations operate parallel tracks under shared oversight. The human track focuses on identity, behavior, and context, supported by empathy, due process, and proportional response. The AI track focuses on alignment, scope, and traceability, supported by technical guardrails, logging, and review. Reporting both tracks together gives leadership one narrative about trust across people and machines.

Program Governance and Adaptation

Governance sustains momentum through funding cycles and organizational change, providing the structure and accountability that keep the program credible over time. Adaptation ensures the program remains fit for purpose as business priorities evolve, technology advances, and regulatory landscapes shift. Together they provide continuity on ordinary days and resilience when conditions change.

Many organizations use a cross-functional steering group to set the rhythm. Security, HR, legal, and IT teams and a senior business sponsor meet on a predictable cadence. Weekly sessions often focus on active cases and key metrics. A monthly forum examines root causes and the control backlog. Each quarter, the conversation widens to the board, covering trends, notable exceptions, and the investments that follow.

Board oversight is easier when defensibility is clear. Programs that publish what is monitored and why, limit access by role, record approvals for identity unmasking, and preserve evidence with chain of custody tend to earn trust. Investment aligns more readily when outcomes are expressed in terms customers, regulators, and auditors recognize: customer impact, regulatory exposure, operational downtime, and brand risk.

Evidence and auditability give the work a stable foundation. Runbooks carry a version, a named owner, and a last review date. A privacy and civil liberties statement remains current and visible. Oversight reviews and self-assessments occur on a regular schedule. An exception register and risk acceptance log make trade-offs explicit rather than implicit.

Treating the program as a living system keeps it aligned with reality. Many teams run a quarterly assurance loop that includes scenario tabletops, control validations, detection sampling, refresh of role and data baselines, and remediation items tracked to closure. Regulatory changes are mapped to specific controls and metrics, with evidence stored in an auditable repository and change notes that set revised targets. As AI becomes embedded in daily work, an agent registry often records ownership and scope attestations, model and version hash, log endpoints, results of kill switch tests, and periodic review of prompts, decisions, and memory events.

Measuring Impact

Metrics turn intent into accountability by showing whether controls are working as designed. The most useful scorecards track two dimensions at once. Leading indicators confirm that guardrails are in place before something goes wrong. Lagging indicators show how effectively the program responded when incidents occurred. Because many exfiltration attempts now pivot through third-party SaaS, it is important to include measures that reflect this path.

Here are some example leading indicators:

- Share of identities under least privilege or with just-in-time elevation
- Proportion of contractors in controlled or virtual desktop environments
- Percent of AI agents registered with owners, scopes, and log endpoints
- Fraction of crown jewel data mapped, tagged, and covered by lineage tracking
- Mean time from risky action to first policy-aware warning
- Share of uploads of protected data to unsanctioned SaaS blocked at time of action

Here are some example lagging indicators:

- Mean time to detect and mean time to contain, reported separately for human and AI cases
- Incident rate per thousand users or agents
- SaaS pivot incident rate and mean time to contain from first risky upload to token revoke and data quarantine
- Top recurring root causes and post-remediation recurrence rate
- False positive rate by detector and by business unit
- Employee perception of fairness and average grievance resolution time

Many teams publish a quarterly scorecard from the central dashboard and use it to fund what works and adjust what does not.

Getting Started: First 90 Days

Before diving into detailed operating models, teams often ask a simple question: "*Where do we start?*". A ninety-day horizon, shown in Table 4-1, provides enough space to build momentum without attempting to change everything at once.

Table 4-1. First 90 days: foundational actions[a]

Visibility	Access	Third parties and AI	Equip and adjust
Identify crown jewels, map access paths, set baselines, and route alerts centrally.	Retire dormant accounts, shorten token lifetimes, strengthen offboarding, and apply just-in-time elevation.	Place high-value systems behind controlled access; register AI assistants and agents with owners, scopes, logs, and a kill switch.	Provide role-based training; run early after-action reviews; refine baselines and playbooks.

[a] Note: This is not a deadline, but a planning compass teams use to shape early program momentum.

Conclusion: From Risk to Trust

An insider risk program is not about surveilling people; it is about enabling productive work without unnecessary exposure. With clarity, fairness, and discipline, investigations become less frequent, responses faster, and trust stronger.

Programs that govern both people and AI-driven systems under a single narrative are better positioned to adapt as the landscape shifts. By aligning insider risk management with the organization's broader risk appetite, programs maintain the balance between protection and productivity. The path forward is steady practice over improvisation, program structure over tool activity, and trust that is earned through accountability.

References

- Arctic Wolf. *The State of Cybersecurity: 2024 Trends Report*. Arctic Wolf Networks Inc., 2024. *https://oreil.ly/h5Vck*.
- Bargury, Michael. "Your Copilot Is My Insider." Zenity Labs, presented at the RSAC 2025 Conference, April 28–May 1, 2025. *https://oreil.ly/9kFJW*.
- Caralli, Richard A., et al. *CERT® Resilience Management Model, Version 1.2.* Carnegie Mellon University, February 2016.
- Cybersecurity and Infrastructure Security Agency (CISA). *Insider Threat Mitigation Guide.* US Department of Homeland Security, November 2020. *https://oreil.ly/rJUm0*.
- Dekker, Nick. *Insider Threat Statistics: Malicious Intent or Ignorance?* Eftsure, May 30, 2025. *https://oreil.ly/A7PQU*.
- IBM Security. *Cost of a Data Breach Report 2025*. IBM Corporation, 2025. *https://oreil.ly/H1xwM*.
- Kapoor, Rahul, and Charlotte Cavendish. *CrowdStrike 2025 Global Threat Report*. CrowdStrike, August 2025. *https://oreil.ly/uXw29*.
- Lynch, Aengus, et al. "Agentic Misalignment: How LLMs Could Be an Insider Threat." Anthropic Research, June 20, 2025. *https://oreil.ly/ZwOWf.*
- Ponemon Institute. *Cost of Insider Risks Global Report 2025*. Ponemon Institute, 2025. *https://oreil.ly/o-9MA*.
- Scroxton, Alex. "CISOs Spending More on Insider Risk." *Computer Weekly,* February 26, 2025.

- Singh, Kanishka. "Barclays Being Probed by UK Privacy Watchdog on Accusations of Spying on Staff." Reuters, updated August 10, 2020. *https://oreil.ly/s51HQ*.
- Verizon Business. *2025 Data Breach Investigations Report*. Verizon, 2025. *https://oreil.ly/Y0fVh*.

About the Author

Reet Kaur is an accomplished cybersecurity executive, Fractional Chief Information Security Officer (CISO), and AI risk strategist who helps organizations align security with business goals across both traditional technology and emerging innovations. As the founder of Sekaurity, she provides full stack security leadership from strategy and governance through operational execution, alongside practical AI governance and quantum risk readiness.

With over 24 years of security and IT leadership experience, including Fortune 100 environments, Reet helps small and mid-sized businesses, public sector institutions, and high-growth startups build and mature security programs through risk-based advisory, governance, and transformation.