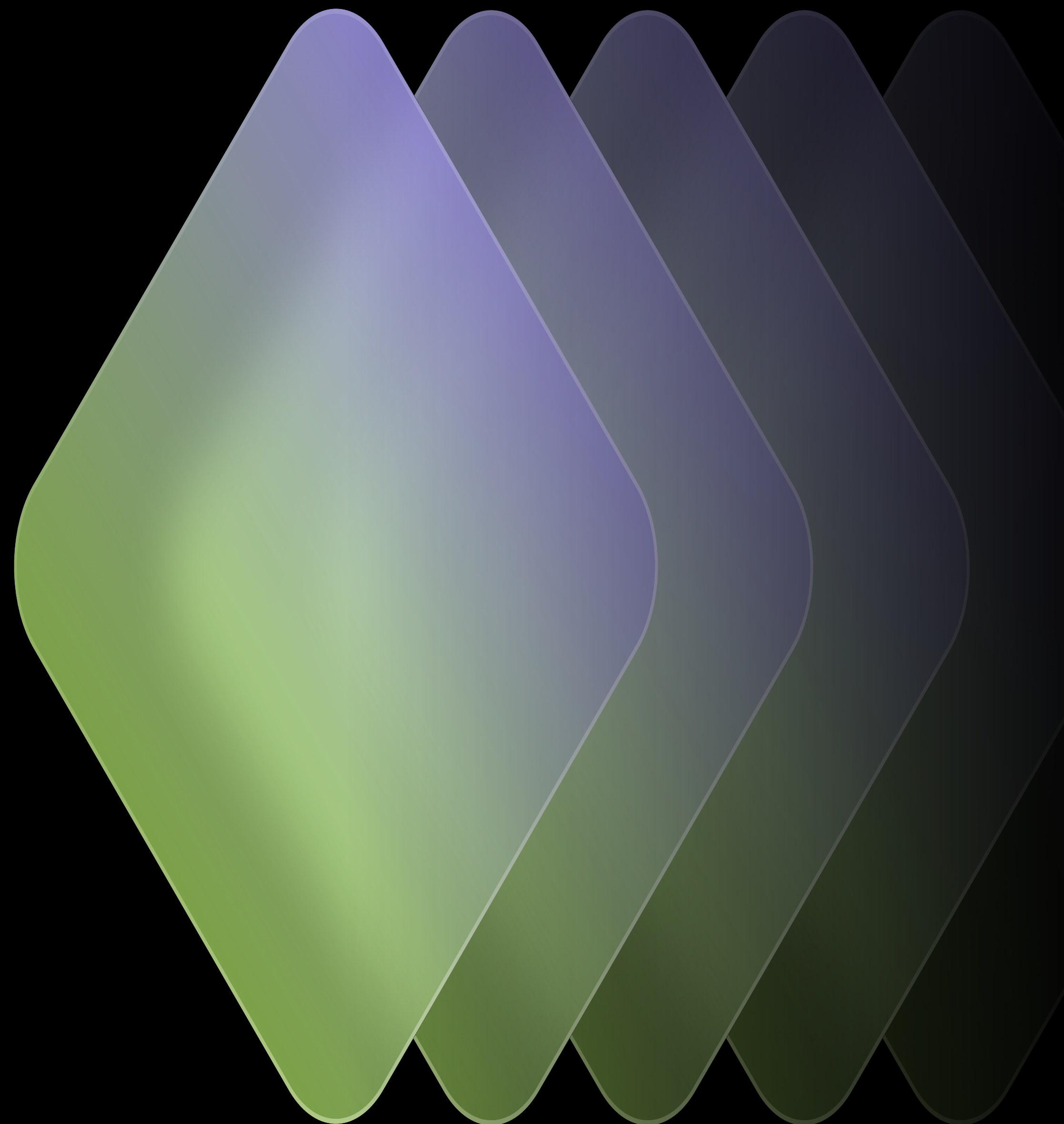


Guide to Deploying AI You Can Actually Trust

2026



We asked IT Directors, Heads of IT, and CIOs at mid-market and enterprise tech companies about their experience and concerns implementing AI. The same worries kept surfacing: misleading output, messy data, security risks, and where human verification fits in.

In this guide, we uncover what stands behind each of these concerns and ways to adopt AI safely. Jump to the section that matters most to you, or read straight through.

Contents

- 01 Weak Access Control and Information Leakage 4
- 02 Potential Security Risks 7
- 03 Misleading Output and Hallucinations 10
- 04 Low Data Readiness 13
- 05 Automation vs. Human-in-the-Loop Balance 16
- 06 Verification Tax Trap 19

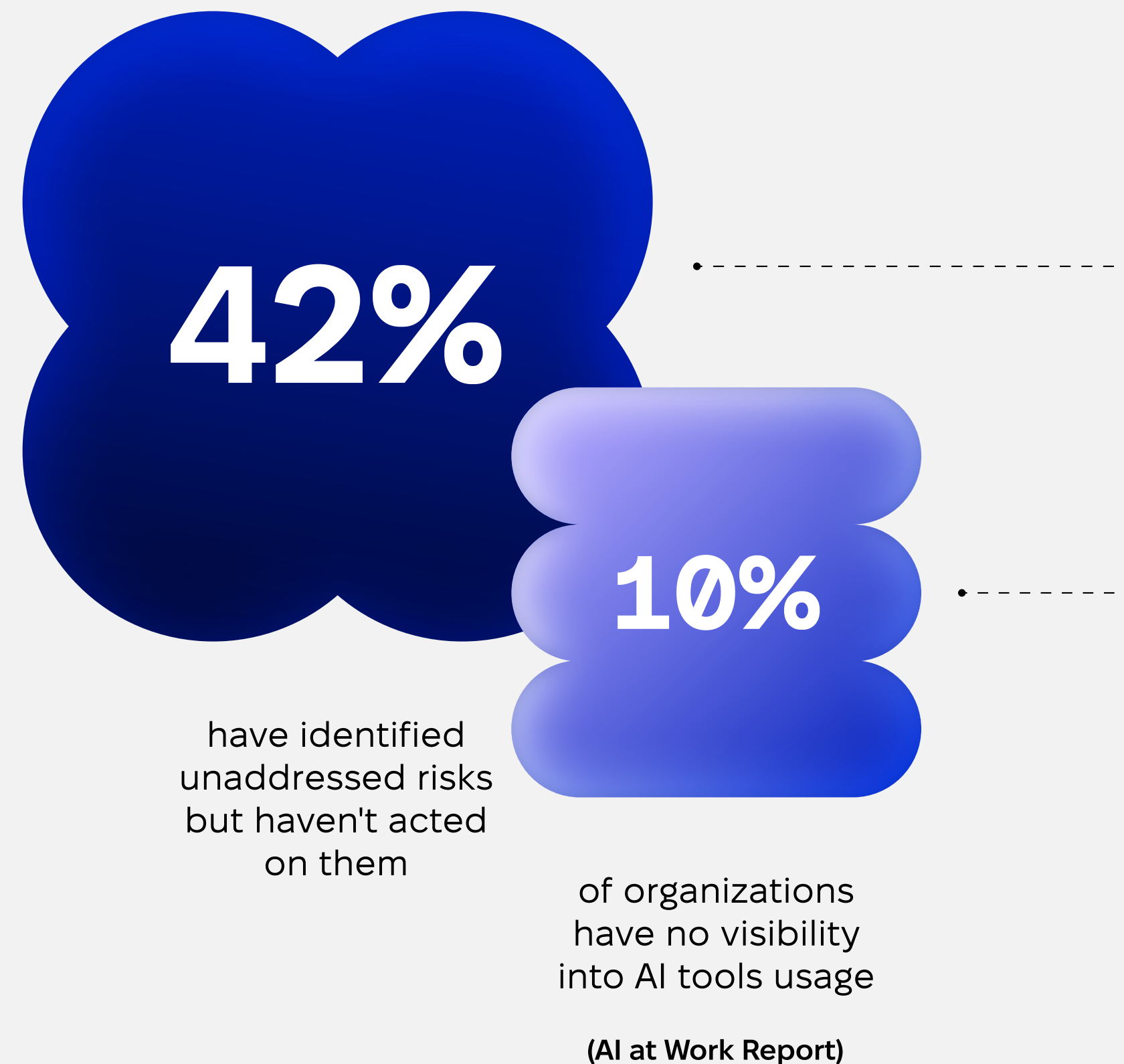
01

Weak Access Control and Information Leakage

"Right now we are still stuck in policy and governance phase. We have teams that want to use AI, but due to data sensitivity and lack of IT support, we struggle to allow the use of AI. I'm still waiting to get my AI policy pushed through Legal."

— Kent, IT Manager / Information Security Officer, Enterprise

Weak Access Control and Information Leakage



What's behind it

AI tools get adopted faster than governance policies. By the time IT catches up, the tool is embedded in daily workflows, already connecting to internal documents that may contain sensitive data. According to IBM's Cost of a Data Breach Report, one in five organizations has faced a breach due to shadow AI, yet only 37% have policies to manage or detect it.

The core gap is permission awareness. Most AI tools retrieve documents based on semantic similarity, not on whether the user is actually authorized to see them. Exposure can happen through legitimate usage, like an employee asking a regular question and accidentally getting access to confidential information.

Risks like that delay AI implementation. Companies often get stuck between the need to optimize and the reality, in which managers evaluate AI tools for months without moving further.

1. Implement RBAC retrieval

For confidential data, reveal only the information that the user is already authorized to see. You can do this through user-level access tokens. Rather than indexing an entire knowledge base upfront, the system authenticates each request as that specific user and retrieves only what they are permitted to access. Since sensitive content is never stored in the index, the system cannot leak data it never held.

2. Control data flows, not tools

Decide which data an AI tool is allowed to touch based on data class and permission scope, not the product name or vendor claims. Build that rule once, apply it to every tool. There's no point in tracking every new feature each vendor ships; you'll always be behind.

3. Move from static permissions to continuous evaluation

Set behavioral baselines and monitor patterns like access frequency, data types, and request volume. Trigger verification or automatic session termination when something unusual happens. While static permissions tell you who was allowed access at setup, behavioral monitoring tells you what's actually happening.

4. Treat the browser as a separate control surface

Enforce browser policies on managed devices, including extension allowlists, approved AI tools, and data classification rules, to manage what employees can upload. Browser extensions are a common and underestimated leakage point. Employees unknowingly upload sensitive information, either to speed up work or test a new tool.

5. Have an AI governance board

Make sure security, legal, IT, and relevant business units have shared visibility and accountability over AI implementation. A board creates a consistent decision-making process for evaluating new tools, developing policies, and handling exceptions, which particularly matters if you handle sensitive customer data and need to achieve regulatory compliance. KPMG's AI governance principles are a useful reference for boards getting started.

02

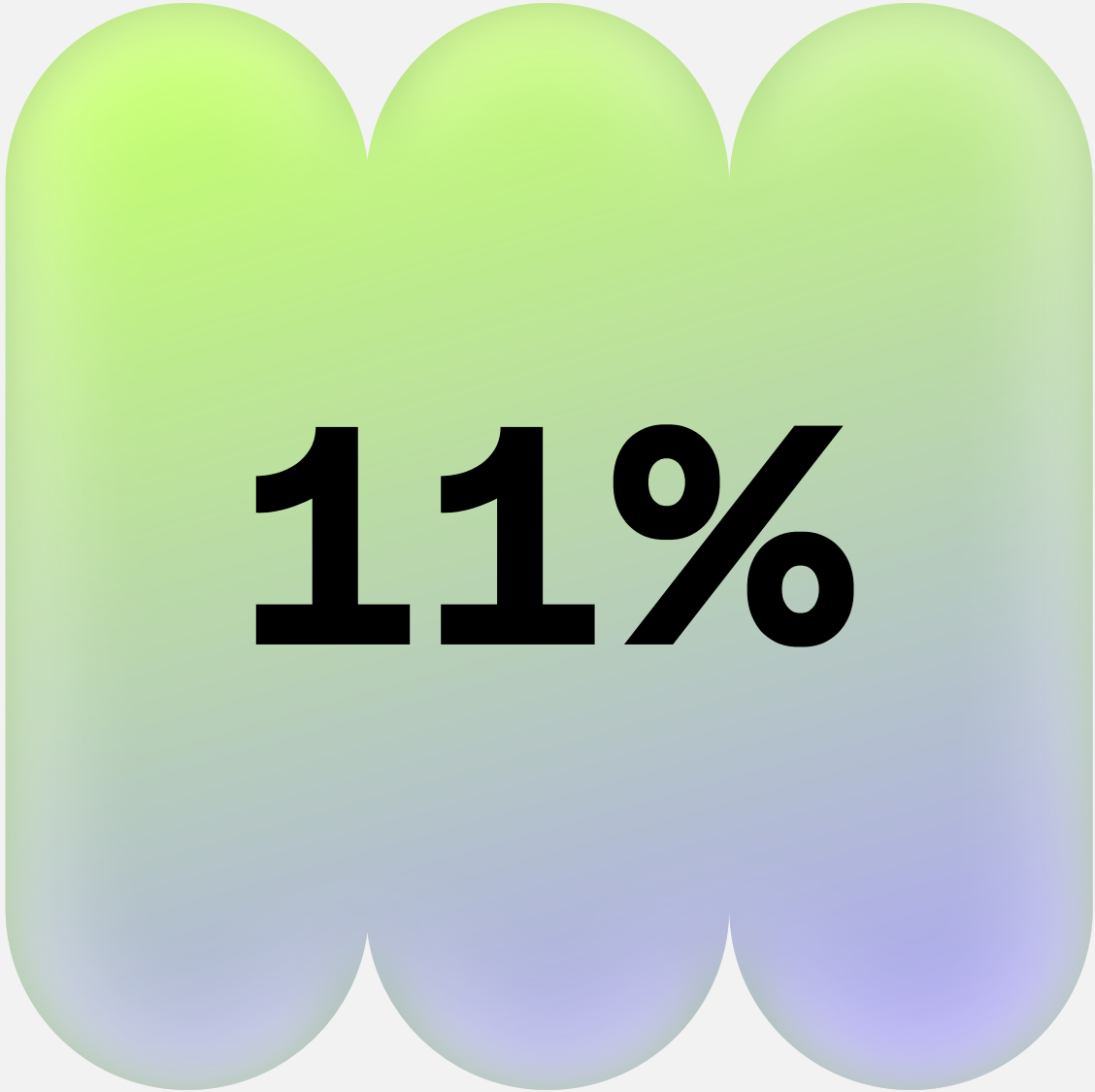
Potential security risks



"No data used as AI training — that's a non-negotiable.
And anything we bring in has to pass our security screen."

— David, Director of Infrastructure, Data-economy product company

Potential Security Risks



11%

of organizations have
experienced a security
incident or near-miss.

(AI at Work Report)

✦ What's behind it

The security risks of your AI vendor are your security risks. The problem is usually how you approach implementation and data processing.

Most enterprise AI deployments rely on OAuth tokens, API keys, and service accounts to access internal systems. These typically request broad, long-lived permissions. An AI tool may only need to retrieve a single document, but still get read access to an entire repository. That over-provisioning creates additional security weaknesses.

Training data is a separate security concern. Many cloud-based AI tools include clauses about using customer data to improve their models. And for organizations handling confidential HR, legal, or financial data, it's a big issue.

Finally, the security review you ran when first onboarding the tool may already be outdated. Vendors ship new AI features continuously, each one potentially expanding the attack surface.

1. Have a pre-deployment security checklist

Before connecting any tool to internal systems, clarify its data access, processing, use for training external models, and liability. Audit what each AI tool can actually access versus what it needs to access, and scope it to the minimum required action.

2. Pilot in a sandboxed environment

Run controlled testing before production to see where internal terminology, tooling, or policies result in unexpected behavior. You can make a test set annotated by humans and benchmark your results against it.

3. Audit what has been launched and keep logs

Run a retrospective audit of deployed tools, prioritizing anything with access to HR, legal, financial, or customer data. You should also continuously validate performance and track every tool invocation, parameter passed, and document retrieved. This will make troubleshooting easier if something happens.

4. Assess risks at the design stage

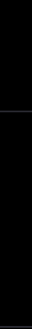
For custom AI development, run structured risk assessments early, including FMEA and ISO 31000 risk management, to detect biases, privacy leaks, and security hazards. For third-party solutions, run vendor due diligence and procurement governance. Cover use case risk classification, AI system ownership, customer data handling, and system maintenance, among other things.

5. Map AI data access to compliance

Every AI tool that touches regulated data needs documented access controls, retention policies, and an audit trail that can withstand compliance review.

03

Misleading Output and Hallucinations



"There are always a lot of false positives and hallucinations. You cannot just use it as is."

— Arun, IT Director, ~600-person media firm

Misleading Output and Hallucinations



2%

trust AI outputs enough
to act on most of them
without review.

(AI at Work Report)



What's behind it

AI models work by predicting the most probable next token based on patterns learned during training. Not every model can tell the difference between current, accurate information and outdated content. As a result, misinformation risks go up.

Most enterprise AI tools pull data using retrieval-augmented generation (RAG) governed by semantic similarity, not recency or authority. If your documents aren't tagged with age, version status, and supersession flags, the retrieval layer cannot filter for accuracy.

Another problem is that when no data exists, many models don't return an error. An employee asking about an undocumented change may get a response that reflects the previous plan or industry norms. In many tools, there is no signal indicating that the model is operating outside of its reliable knowledge.

1. Audit the knowledge base and add metadata

Date every source document, mark superseded content, and assign an owner responsible for feeding the model with clean, current documentation. Adding metadata to the retrieval layer also helps the model to return the right chunk, not just the most topically similar one. The research on utilizing metadata for RAG shows that company and year fields matter most, while section titles are secondary.

2. Limit what the model handles autonomously

Set escalation rules in tools like Slack and Confluence for categories that may require a human look, like access permissions, compensation, and legal compliance. If a domain keeps generating flagged outputs, the documentation is almost certainly the problem and needs fixing.

"We use a two-stage retrieval process. After the initial search and similarity-threshold filtering, a reranker reevaluates the results based on actual relevance and removes weaker matches."

– Yevhenii Mostipan, Senior Software Engineer at Leebry.

3. Audit what has been launched and keep logs

Run a retrospective audit of deployed tools, prioritizing anything with access to HR, legal, financial, or customer data. You should also continuously validate performance and track every tool invocation, parameter passed, and document retrieved. This will make troubleshooting easier if something happens.

4. Require citations or express uncertainty

Implement structured response templates that make the model cite a specific document for every output and flag when none exists. When the knowledge base comes up short, the model should either narrow the query and retry or tell the user it lacks context to answer.

5. Start with low-stakes workflows

Automate basic, repetitive tasks first. Expand autonomous handling only after accuracy patterns are well-established and documented. Moving step-by-step also matters for smooth adoption. Poor early results stick, making people reluctant to use AI tools.

If you need help differentiating low- and high-risk workflows, we have a tool based on the NIST AI Risk Management Framework to guide you.

04

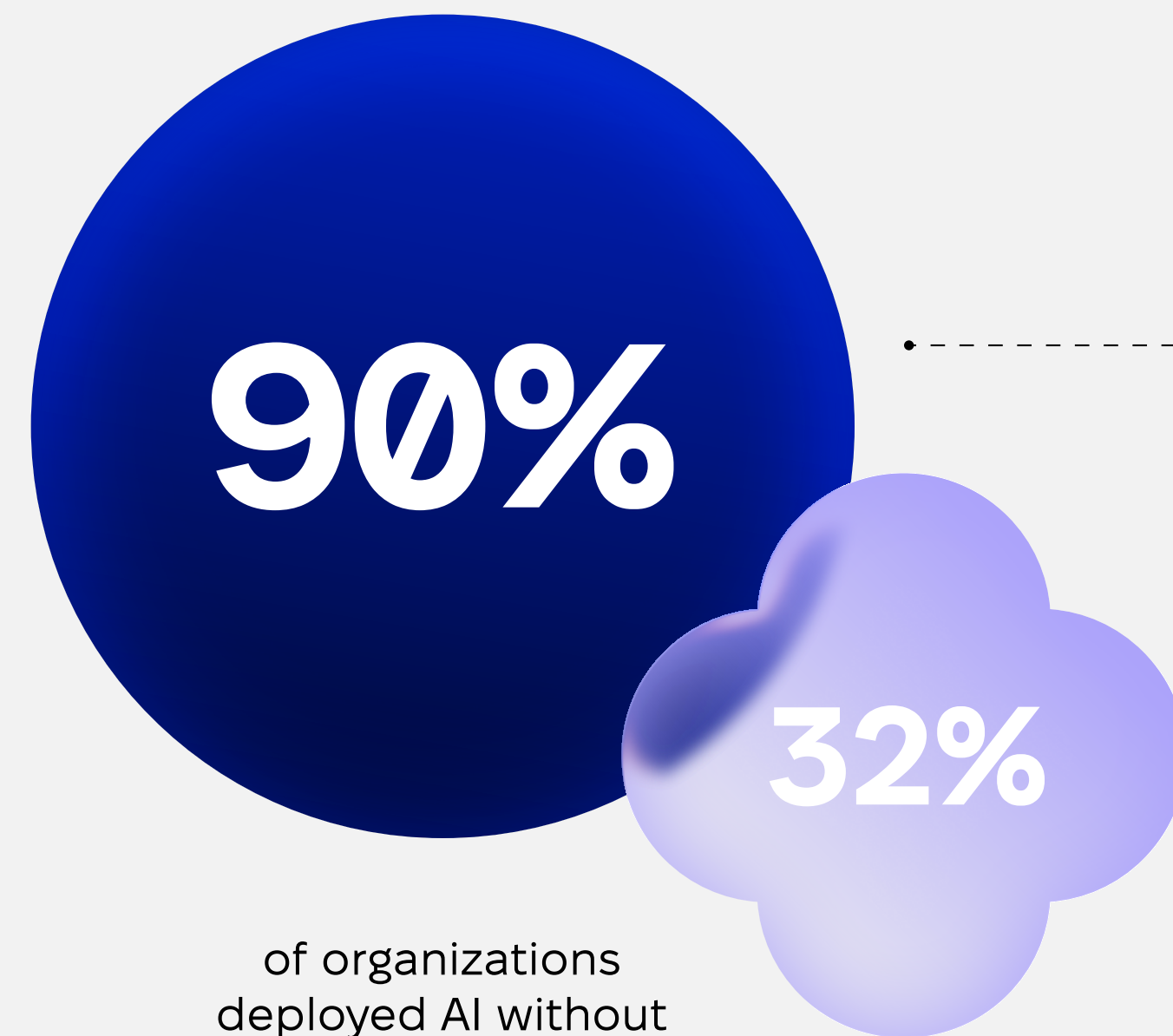
Low Data Readiness



"Data readiness is the elephant in the room. Leaders want AI, but their data is messy, scattered, and poorly governed. The gap between vendor promises and reality is huge."

— Zuri, Director of Data and Tech Transformation, Financial services

Low Data Readiness



of organizations deployed AI without extensively auditing internal knowledge.

estimate their knowledge base is less than 50% accurate.

(AI at Work Report)

What's behind it

AI tools are layered on top of years of accumulated documentation, often without cleanup. The results in our AI at Work report are concerning. **26%** of IT leaders estimate their knowledge base as **25–50%** accurate, and **6%** put it below **25%**.

Stale content is part of the problem. AI performs reliably only when its source data is standardized with consistent taxonomies, naming, and formatting. Most organizations haven't done that work before deployment, which compounds the accuracy problem.

Another reason behind low data quality is a lack of governance and data ownership across multiple Confluence spaces, Slack threads, and Google Drive folders. Teams add content, rename processes, and update policies without clear rules on accuracy, review cycles, and change approval. Organizations end up with lots of shadow data, which tends to come up after AI implementation.

1. Catalog and label data

Ask teams to catalog what data exists and differentiate between current and superseded. Frequently updated records like policies, compliance documentation, regulatory guidance, and internal knowledge bases are usually considered stale after 12 months.

"Documents should be well-structured, with clear headings and descriptions. Try to include fewer complex tables, dynamic content, links, and images, as they complicate things."

– Yevhenii Mostipan, Senior Software Engineer at Leeby.

2. Assign ownership to every knowledge area

Make sure each L1 support procedure, HR workflow, and other operations you automate with AI have a named person responsible for accuracy and a defined review cadence. Consistent naming conventions, taxonomies, and formatting also help the model interpret and connect information across documents.

3. Build continuous data maintenance

Make deduplication, versioning, and content expiration ongoing workflows. Some AI tools can help, flagging outdated content, excluding stale data from indexing, and catching contradictions across documents.

4. Start with clean data

Focus on domains that have well-maintained documentation. Pick a pilot team or department and use clean data for automation. Add a simple feedback mechanism (thumbs up/down) and rate outputs for two weeks to measure quality.

5. Manage information ownership directly

Assign document owners and review cycles inside an AI platform. Owners should be notified when their document is flagged as outdated/disputed and have a clear path to approve or reject edits.

05

Automation vs. Human-in-the-Loop Balance

"How can AI automate IT repetitive processes without giving it too much control or access that ends up doing more harm than good?"

— Grace, IT Director, IT services

Automation vs. Human-in-the-Loop Balance



61%

of IT leaders report a gap between leadership expectations and what can be delivered.

(AI at Work Report)

What's behind it

The boundary between "AI suggests" and "AI acts" shifts too easily. It often happens without a formal decision and careful testing. 61% of IT leaders we've surveyed report a gap between what leadership expects AI to deliver and what their team can operate safely.

It's hard to determine where human intervention is still critical because the answer depends on more than just whether AI can perform a task. Organizations have to build workflows individually, taking into account the required accuracy, risks, accountability, ethics, and costs.

AI-related pressure often translates into expanded automation scope and people getting removed from critical processes when they shouldn't. A wrong answer in draft mode costs a few seconds of review time. The same wrong answer sent autonomously to an employee, a customer, or a compliance system costs significantly more.

1. Define handoff criteria and triggers

Decide explicitly which ticket categories AI can resolve autonomously, which require human review, and what triggers an escalation (e.g., low confidence scores, negative sentiment, repeated clarification requests, personal data, or access permissions). High-confidence outputs in established ticket categories can reasonably skip human review. Anything that touches access, compliance, or personal data needs a human check. Also, flag outputs with no cited source for automatic human review.

"High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use."

– EU AI Act.

2. Start in review mode for testing

Make action autonomous only after it shows a specific accuracy threshold over a defined period. No acceptance threshold is objectively "correct". Accuracy is a normative choice that depends on the context. Setting it higher reduces harm but may block useful systems; setting it lower allows deployment but tolerates more errors. As a rough guide, 70–85% accuracy may still be useful for low-risk workflows if it saves time. Medium-risk workflows generally require 85–95% accuracy if errors are cheap and harmless. When risks are high, human oversight is essential.

3. Keep audit trails and compliance controls

Log AI-assisted decisions, including what was generated, related sources, and who reviewed the output. It helps investigate when something goes wrong and demonstrate compliance.

4. Use automated validation before human review

Run deterministic checks first to confirm AI output follows hard-coded logic, business constraints, and formatting rules. You can also use an LLM as a judge to score the primary model's output against a written rubric, in place of or alongside human raters.

06

Verification Tax Trap



"It still needs that expert review, and those experts are already tied down with 50 other things."

— Arin, VP of Technology Strategy, Architecture & Information Security,
SaaS company

1. Focus on use cases where verification is fast

Prioritize repetitive operations that are verifiable in seconds. AI can deflect repetitive tickets, answer policy questions, handle onboarding/offboarding at scale, or route access requests. Verification should not take as much time as doing it manually.

"Not just understand the "pain" AI solution will solve, but prepare 2-3 use cases and write scenarios. How will your colleagues use it? Is it convenient and aligned with the other toolstacks they use daily?"

- Lina Klymovych, Customer Success Lead at Leebry.

2. Connect AI to real user and system context

Reduce edge cases that require human review with personalized, context-aware AI responses. Once connected to identity data, ticket history, and system state, AI tools have broader information about the user requesting something and can provide more relevant output, reducing the need for human review.

3. Know end-to-end task time

Measure net minutes per task in every pilot, not gross savings. You should compare AI time plus verification time against the manual baseline to ensure the productivity gain is real.

4. Track the real impact

Check whether the current AI setup solves the underlying problem by tracking resolution rates, error rates, reopen or re-escalation rates, and user satisfaction by workflow category. Benchmark the workflow before AI adoption, so you have something to compare against.

"Monitor after-implementation changes: is the tool really helpful or it's just nice to say you have AI in place? Prepare metrics to track BEFORE implementing."

- Lina Klymovych, Customer Success Lead at Leebry.

5. Scope verification before deployment

Define human oversight touchpoints, what must be checked, and how long verification should take. It helps see potential constraints and whether you have the required capacity from the start.

Implementation Takeaways

01

Clean data before the implementation. Include it in the project scope and dedicate 2x more time to it than you originally think.

02

Create a pilot group within your organization. People who are ready to dedicate time to reviewing, testing, and controlling the quality of responses at the initial stages.

03

Give AI tools only the permissions the requesting user already has. Authenticate users each time they retrieve information for extra security.

04

Decide what AI can do alone and what needs human oversight. Anything touching compliance, access, money, or personal data gets a mandatory human check.

05

Measure net time, not gross time saved. Always compare AI time plus verification time against the old manual process to see the real impact.

06

Pilot in low-stakes, easily verified workflows. Expand AI use and autonomy only after testing the accuracy over time.

How The Guide Was Made

This guide is based on industry research and our AI at Work report gathering insights from 100 participants holding Director-level or higher IT roles. We also talked to people adopting AI directly to hear about their pains and worries.

The next stage was asking our engineering, security, and legal teams about possible solutions. That’s how we accumulated expert advice on how to make AI secure and manageable, reduce risks, and find a balance between automation and human in the loop.

Authors:

Lina Klymovych, Customer Success Lead
Yevhenii Mostipan, Senior Software Engineer

Sources:

- [AI at Work by Leebry](#)
- [OWASP Top 10 for Agentic Applications for 2026](#)
- [Cost of a Data Breach Report 2025](#)
- [AI Governance Principles for Boards](#)
- [Artificial Intelligence Risk Management Framework \(AI RMF 1.0\)](#)
- [EU AI Act](#)
- [Is your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act](#)
- [Utilizing Metadata for Better Retrieval-Augmented Generation](#)
- [LLM as a Judge: The Complete Guide](#)

About Leebry

Leebry is a Work AI platform designed to help organizations make better use of the knowledge and tools they already have. By connecting systems like Confluence, Slack, Jira, Okta, and HRIS, it provides secure, citation-backed answers and enables everyday workflows through a single interface.

Leebry is built by MacPaw, known for products like CleanMyMac, Setapp, and ClearVPN, as part of its move into enterprise AI.

[Learn more at leebry.com](https://leebry.com) ↗

