1

Context-driven architectures will be everything

The success of enterprise AI initiatives is reliant on enabling AI agents to effectively and securely access the most relevant data and tools, empowering them to deliver unique and differentiated products and services to customers and clients. As end-to-end automation transforms the software development lifecycle to support the volume of code generated by AI tools, developers will focus less on manual coding and more on architecting context-rich applications, using AI tools and context engineering techniques.

Physical AI

Integration of artificial intelligence with real-world environments, enabling smart devices and robots to autonomously perceive, reason, and interact through sensors and edge devices.

Physical AI represents the convergence of artificial intelligence and physical hardware systems, empowering intelligent agents to perceive, reason and interact with the real world through sensors, actuators and edge devices. This integration enables robots, automated machines and smart systems to act, learn and adapt within physical environments, effectively bridging the digital and physical realms.

Physical AI models are trained on physical interactions, spatial relationships and the laws of physics themselves. Using advanced simulation environments, these systems can experience millions of scenarios in virtual replicas of the real world, learning how objects behave, how forces interact, and how to manipulate physical space. Through techniques like reinforcement learning, sim-to-real transfer, and synthetic data generation, models develop an intuitive understanding of physics, geometry and causality that enables them to operate effectively in dynamic, unstructured environments. Once trained in simulation, these models are deployed to physical hardware where they continue learning from real-world feedback, constantly refining their understanding of physical reality.

The applications span manufacturing, logistics, and robotics. In warehouses, robotic systems trained on millions of simulated pick-and-place scenarios can handle diverse objects they've never encountered, adapting grip strength and approach angles based on visual and tactile feedback. On manufacturing floors, AI-powered quality inspection systems learn to detect defects across varying lighting conditions and product variations by training on synthetic datasets that mirror real production environments. Autonomous mobile robots navigate complex facilities by combining pre-trained spatial reasoning models with real-time sensor fusion, learning optimal paths while avoiding dynamic obstacles. Robotic arms in assembly lines learn intricate tasks through demonstration and practice in simulation, then transfer that knowledge to physical production with minimal real-world fine-tuning.

The trend toward AI-driven physical automation is being propelled by several converging technological and economic factors. Simulation breakthroughs have enabled advanced physics engines and digital twin platforms to create photorealistic, physics-accurate virtual training environments at scale, while improved sim-to-real transfer techniques have significantly reduced the reality gap by allowing knowledge gained in simulated environments to translate more effectively to physical applications. AI-powered synthetic data generation tools now produce unlimited labeled training data representing diverse physical scenarios, complemented by hardware advances that bring more powerful edge computing chips and sensors capable of real-time AI processing directly to the point of action. The proliferation of IoT devices and sensors has triggered a data explosion, generating vast amounts of physical world data that enables continuous learning and optimization. Meanwhile, the decreasing costs of robotics components, sensors, and compute power have made deployment economically viable across industries. Finally, persistent labor challenges, particularly workforce shortages in manufacturing and logistics, are driving heightened demand for intelligent automation systems capable of adapting to varied and complex tasks.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

McKinsey predicts the physical AI market is projected to reach $370B+ by 2040 driven by enterprise adoption across diverse industry applications such as facilities management, physical security, manufacturing and logistics.[1] This is driven by investments in platforms that deliver measurable ROI through energy savings, labor reduction and predictive maintenance.

Physical AI has emerged with a variety of initial use case applications across multiple domains. Smart building and IoT platforms are deploying AI-powered building management and environmental control systems that optimize facility operations. Spatial intelligence platforms are being developed by companies building AI models that understand three-dimensional physical spaces and enable autonomous navigation through complex environments. Embodied AI and robotics applications are introducing physical robots and autonomous agents capable of performing inspection, delivery, and various facility tasks. Simulation and training platforms are creating synthetic environments specifically designed to train Physical AI systems before real-world deployment. Computer vision and perception technologies are providing visual AI capabilities for monitoring, security, inspection, and safety applications across industries. Finally, edge AI infrastructure is delivering the hardware and software necessary to enable AI processing directly at endpoints without relying on cloud connectivity, ensuring faster response times and greater operational independence.

Physical AI presents several key implications for the enterprise across operational domains. In smart buildings, security cameras are leveraging AI to detect real-time threats and send immediate alerts, while access control systems utilize biometrics for enhanced security, and building management systems autonomously control temperature, lighting, and ventilation based on occupancy patterns and environmental conditions. Autonomous operations are transforming warehouses, manufacturing facilities, and logistics operations through the deployment of AI-powered robots that handle material movement, conduct quality inspections, and manage delivery tasks. Enhanced customer experiences are being realized in retail environments where Physical AI enables sophisticated inventory management, checkout automation, and personalized in-store assistance that adapts to individual shopper needs. Additionally, predictive maintenance capabilities are revolutionizing industrial operations as equipment outfitted with sensors and AI can predict failures before they occur, significantly reducing costly downtime and enabling proactive intervention.

[1] Will embodied AI create robotic coworkers?. McKinsey & Company. (2025, June 30). https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/will-embodied-ai-create-robotic-coworkers.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Private market data insights exchange for the AI economy

Expansion of real-time insights from structured private market data through seamless, consumption-based exchanges, and enabling the trading of data as an asset class.

Traditional private market data providers are known for providing data attributes tied to company entities (e.g., funding, valuation, investors, job postings), brands (e.g., adverse media), transactions (e.g., location, time), people, and beyond. This private data is typically ingested by corporations through an API feed and further used to enrich existing firmographic data. One challenge with this existing approach is that (1) private markets data is often not structured using a universal entity framework and (2) data is often stuck behind walled gardens, or (3) contractually limited to specific use cases and individual user access, thereby being underutilized.

The rapid growth of both unstructured data as well as agentic capabilities, has given rise to Private Market Data Insights Exchanges. These live data exchanges are built on proprietary entity frameworks which help facilitate the buying and selling of transactional data to a broader population of end-users due to less cost barriers. In addition, some exchanges leverage a waterfall enrichment method, where an individual data attribute (e.g., revenue, funding, e-mail contact) is indexed across multiple source data providers and then scored to maximize data quality and coverage. Further, these exchanges have potential to power LLMs and enterprise applications on a consumption based model.

This consumption based model will involve trading data as an asset class, where these exchanges will operate similarly to trading securities. AI agents will communicate with other AI agents to exchange data insights, and further through model context protocol (MCP) capabilities have potential to generate executive level insight oriented tasks in a much more streamlined manner than humans.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

The estimated total addressable market for Private Markets Data is expected to grow at a 14.5% CAGR from $8B in 2024 to $18B by 2030 (BlackRock).[2] This is driven by overall increased demand for access to private markets, desire to enhance client offerings (e.g., deeper insights into alternative investments) and to gain understanding of performance and drivers of returns.

The emergence of Private Market Data Insights Exchanges will drive the following outcomes:

Improvement in Structured Datasets: The rise of unstructured datasets, makes data rationalization difficult. With the use of proprietary entity frameworks, this can be improved by leveraging unique identifiers to aggregate disparate data sources
Improvement in Verifiable Data: Novel techniques such as waterfall enrichment helps triaging of datasets for data completeness and verifiability

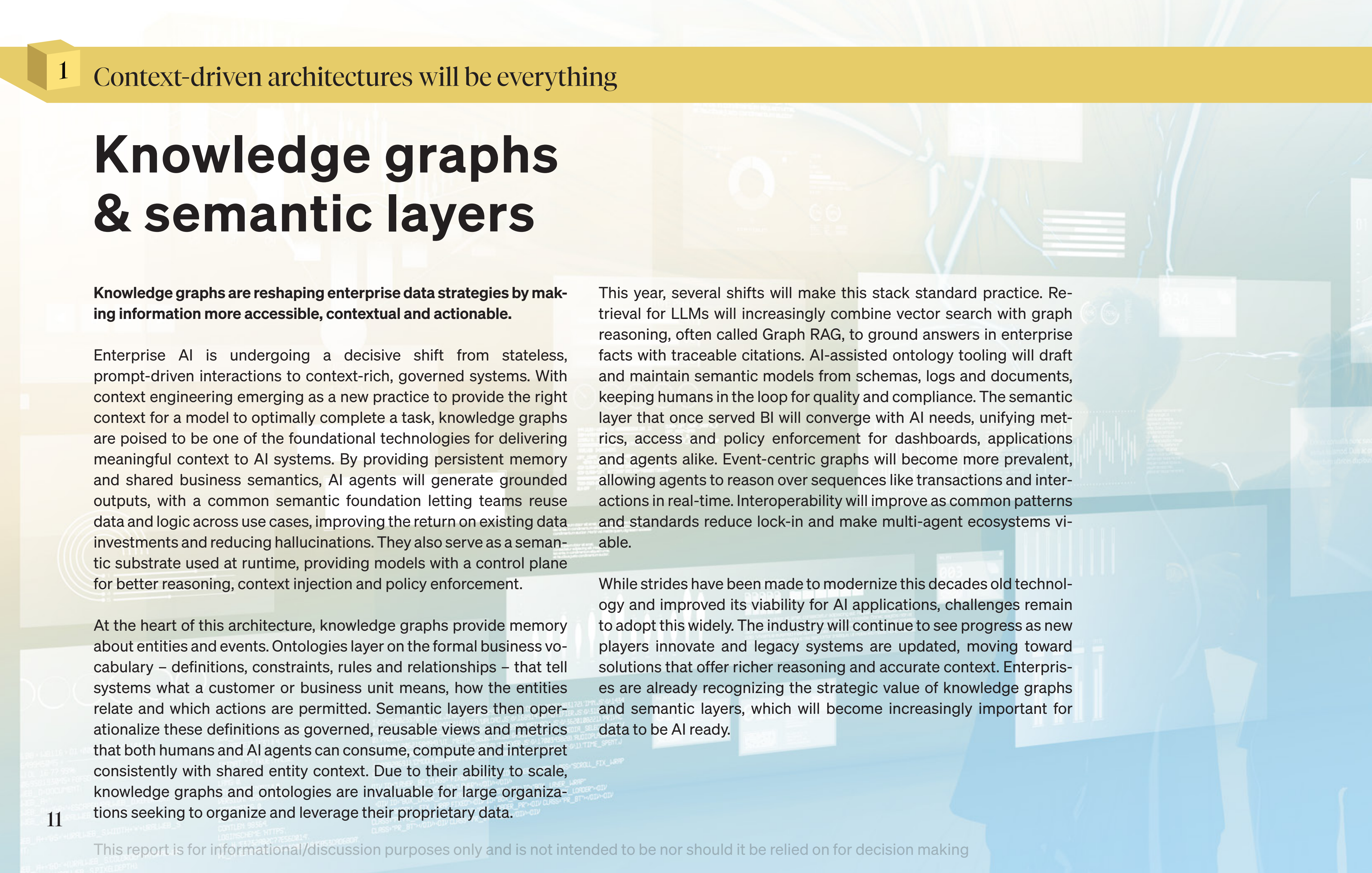
Broader Access to Private Market Data: Emerging data insights exchanges will gain material adoption by natively offering hundreds of primary data sources, which will further be leveraged via API data feeds into homegrown databases in some scenarios this will replace entire usage of traditional market data providers

Actionable Insights & Recommendations: Ability to automate the analysis of raw data, and segment data with the assistance of AI to generate critical insights and messaging to desired audiences, such as internal executives or external sales targets

Accelerated Adoption of AI Research Agents: Leveraging AI to assist in daily tasks (e.g., research and querying, automated e-mail reach out campaigns for GTM teams) and stay ahead of emerging market signals (e.g., headcount changes, social listening, product reviews, adverse media)

Consumption-Based Model / Data As An Asset Class: Traditional market data providers may benefit from a revenue share model with the emergence of private market data insights exchange; however traditional providers may be negatively impacted by the cannibalization of their core user base.

[2] Will embodied AI create robotic coworkers?. McKinsey & Company. (2025, June 30). https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/will-embodied-ai-create-robotic-coworkers



Knowledge graphs & semantic layers

Knowledge graphs are reshaping enterprise data strategies by making information more accessible, contextual and actionable.

Enterprise AI is undergoing a decisive shift from stateless, prompt-driven interactions to context-rich, governed systems. With context engineering emerging as a new practice to provide the right context for a model to optimally complete a task, knowledge graphs are poised to be one of the foundational technologies for delivering meaningful context to AI systems. By providing persistent memory and shared business semantics, AI agents will generate grounded outputs, with a common semantic foundation letting teams reuse data and logic across use cases, improving the return on existing data investments and reducing hallucinations. They also serve as a semantic substrate used at runtime, providing models with a control plane for better reasoning, context injection and policy enforcement.

At the heart of this architecture, knowledge graphs provide memory about entities and events. Ontologies layer on the formal business vocabulary – definitions, constraints, rules and relationships – that tell systems what a customer or business unit means, how the entities relate and which actions are permitted. Semantic layers then operationalize these definitions as governed, reusable views and metrics that both humans and AI agents can consume, compute and interpret consistently with shared entity context. Due to their ability to scale, knowledge graphs and ontologies are invaluable for large organizations seeking to organize and leverage their proprietary data.

This year, several shifts will make this stack standard practice. Retrieval for LLMs will increasingly combine vector search with graph reasoning, often called Graph RAG, to ground answers in enterprise facts with traceable citations. AI-assisted ontology tooling will draft and maintain semantic models from schemas, logs and documents, keeping humans in the loop for quality and compliance. The semantic layer that once served BI will converge with AI needs, unifying metrics, access and policy enforcement for dashboards, applications and agents alike. Event-centric graphs will become more prevalent, allowing agents to reason over sequences like transactions and interactions in real-time. Interoperability will improve as common patterns and standards reduce lock-in and make multi-agent ecosystems viable.

While strides have been made to modernize this decades old technology and improved its viability for AI applications, challenges remain to adopt this widely. The industry will continue to see progress as new players innovate and legacy systems are updated, moving toward solutions that offer richer reasoning and accurate context. Enterprises are already recognizing the strategic value of knowledge graphs and semantic layers, which will become increasingly important for data to be AI ready.

Market and industry perspectives

The knowledge graph market has been around for years. Early adopters in this space either focused on ontology or knowledge graphs. However, we have seen convergence within in the market as the space has grown.

Graph companies have gained adoption with large enterprises due to scalability and fast multi-hop query speed. However, these companies store data as code, freezing the code at a point in time rather than having the data engine figure out connections and conclusions itself.

A few companies recently have focused more on the ontology layer, equipping platforms with ontology models and virtualization engines that are data and application agnostic, meaning they can connect to a number of data sources and tools, including Agent SDKs.

Larger enterprises focus on end-to-end knowledge graph suites that sit on top of an enterprise's data lakehouse, enabling ontologies, applications and agentic frameworks in one platform for AI use cases.

Newer startups in the market are aiming to provide LLMs memory and knowledge. One approach is through a hybrid datastore architecture combining graph, vector, and key-value stores. Another is through building a temporal knowledge graph for AI agents building and updating its graph from not only structured business data, but also from user interactions (chat, unstructured text), tracking when data becomes invalid or changes over time.

Snowflake spearheading the Open Semantic Interchange (OSI) initiative, standards for how entities, metrics and policies are described and exchanged across tools, already indicates how the industry is converging towards governed, semantic first AI architectures.

Data formats for AI

Large-scale, unstructured multimodal data driving the evolution of data infrastructure to better support agentic workloads.

For over a decade, file formats like Parquet (2013, out of Cloudera), Avro (2009, out of Hadoop), and ORC (2013, out of Hortonworks) have served as the backbone of analytical processing. These formats sit on top of object storage (e.g., S3) and are designed to optimize data for analytics.

Sitting above file formats are table formats, which have been a highly competitive battleground in recent years. Open-source Iceberg and Delta have sparked intense industry debate, with Iceberg ultimately becoming the most widely adopted, with ongoing efforts to make both formats interoperable. Table formats add a metadata layer atop file formats, organizing raw files into database-like tables. This lets users store all their data—even structured data typically designed for warehouses—in open, cost-effective formats, while maintaining warehouse reliability and performance without proprietary copies. Users can leverage any compute engine directly on the data, removing the need for data movement or duplication.

While complete ownership, flexibility and interoperability within this stack have provided significant benefits to enterprises, limitations remain. Technologies such as Parquet and Iceberg have been predominantly catered for structured / tabular data, serving batch business intelligence workloads (e.g., SQL).

With the emergence of GenAI and the forthcoming wave of agentic applications—which are proficient at processing highly unstructured and multimodal data, including documents, images and video—new AI-native data formats, both at the file and table level, are being developed to address the evolving requirements of next-generation AI workloads and fill the void of where Parquet and Iceberg fall short.

Open data formats like Lance provide both a table and file format specifically designed for efficient search and retrieval of highly complex multimodal data at massive scale. Open file formats like Nimble, developed by Meta, address Parquet's limitations in AI training by enabling faster reads and more efficient memory layouts. Vortex has also emerged as a Parquet alternative, optimized for AI-native workloads. While these formats are designed for AI workloads, they also support traditional SQL and data engineering (e.g., Spark) processing.

Collectively, these formats seek to overcome the limitations of Parquet and Iceberg in supporting AI and agentic workloads, positioning users to more effectively leverage high-dimensional multimodal data at scale for AI and agentic applications.

Market and industry perspectives

According to Gartner, unstructured data now accounts for 80 to 90% of all new enterprise data and is growing three times faster than structured data.[3] This shift is driving major changes across the data ecosystem.

Leading data platforms, along with hyperscalers that have widely adopted Parquet and Iceberg, are responding to this emerging ecosystem threat.

To address the rising complexity of managing and leveraging unstructured data, AI-focused companies are beginning to shift formats. Netflix, where Iceberg originally developed, adopted the Lance data format to power its multimodal data lake, which includes video, audio, images, text, and embeddings. Further, GenAI native companies have adopted Lance internally to power diverse workloads.

Given the industry-wide investment in Iceberg, and recognizing the shifting landscape, the Apache Iceberg Community is currently reviewing the File Format API Proposal, which seeks to establish a unified, pluggable interface for integrating new file formats with Iceberg—much like its current support for Parquet, Avro, and ORC.

[3] Gartner. (2025, September 17). Market Guide for Data Security Posture Management. Joerg Fritsch, Brian Lowans, and Andrew Bales. https://www.gartner.com/en/documents/6964866

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Context engineering

Context engineering orchestrates external information and tools around LLMs to deliver consistent, accurate and domain-specific AI results.

In the first wave of generative AI (GenAI) adoption, interactions were primarily “prompt-response” exchanges, where users would submit a question and receive an answer. During this period, prompt engineering emerged as a practice, focused on crafting and refining prompts to elicit accurate and relevant outputs from models. GenAI use cases have since evolved toward multi-step agentic workflows, where agents autonomously gather information, use tools, and reason over results with minimal human intervention. To manage the increasing volume of information that agents generate, while balancing the token constraints of context windows (e.g., amount of information that can be provided to the model), context engineering has emerged to curate the optimal set of tokens for achieving desired outcomes across multiple steps.

In agentic applications, effective performance depends on managing multiple types of context: instructions, knowledge, and feedback from tools. Instructions include system prompts that guide model behavior and procedural memory for storing specific skills or rules related to a task. Knowledge consists of facts and relevant experiences, often accessed through Retrieval Augmented Generation (RAG) workflows using vector databases or knowledge graphs. Increasingly, the industry is adopting “just-in-time” context strategies, allowing agents to retrieve only necessary information at runtime instead of pre-processing all data in advance. Throughout these workflows, agents must also integrate feedback and new information gathered from tool interactions, using this context to inform next steps.

While the promise of larger context windows – where most information could be uploaded for a model to use — sounds promising, challenges remain. Processing tokens at such scale drives up computational costs and increases latency, and models, much like humans, are limited by a finite attention budget. As the number of tokens in the context window grows, a model’s ability to accurately recall information from that context declines – a concept known as context rot. Today, industry best practices emphasize that effective context engineering is about selecting the smallest set of high-signal information that maximize the likelihood of achieving a desired outcome.

To achieve this, new techniques like compression are emerging. When the number of tokens approaches the context window limit, this method is used to summarize the most relevant information or filter out less important details. Another strategy involves giving agents a “scratchpad“ – a dedicated space for note-taking that is stored outside the context window and can be retrieved as needed. Lastly, sub-agent architectures offer a way to manage context more effectively; instead of a single agent maintaining state across an entire workflow, specialized agents can focus on specific tasks. As models continue to advance, the challenge of engineering the right context to achieve desired outcomes over long time horizons will remain central to building more performant agents.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

While the developer and coding space has been the natural early adopter of context engineering given the large investments in developer agents, initial efforts were often tool-specific and fragmented. Community-driven standards have emerged from developer workflows in early 2024, functioning as simple, static “READMEs for agents” embedded directly in repositories, providing baseline instructions for agents navigating a codebase. Similarly, specialized tools have emerged to automate the analysis of complex codebases, allowing AI agents to generate structured documentation and “on-demand encyclopedias” to understand millions of lines of code without manual human onboarding.

However, as the market matured in 2025, the industry is moving beyond developer tools toward holistic, cross-platform standards that aim to solve the context problem for every business domain.

The first and most established of these is the Model Context Protocol (MCP). Launched by Anthropic in November 2024, MCP is now a widely adopted open standard that provides a universal “USB-C port” for AI. It standardizes how models or agents connect to external tools – whether third-party applications or proprietary internal APIs – ensuring secure and consistent data access across different platforms.

Building on this connectivity is the invention of “Skills”, launched by Anthropic in October 2025. While MCP handles the connection, Skills provide the procedural knowledge. At their core, Skills are modular folders containing instructions, scripts and resources for specific tasks. Instead of overwhelming the context window with every possible instruction upfront, the agent dynamically “discovers” and loads a Skill only when it becomes relevant to the task at hand.

These two technologies are deeply complementary: MCP facilitates the secure “plumbing” to a tool, while Skills provide the domain expertise to transform that raw access into reliable outcomes. Following Anthropic’s December 2025 release of the Skills open standard, the paradigm has seen cross-industry adoption, most notably by OpenAI within ChatGPT and its Codex developer products.

Reinforcement learning environments

Enabling agents to tackle complex, real-world tasks through goal-driven training, interactions with tools, realistic simulations and outcome-based feedback.

In last year's trends report, we highlighted the rise of reasoning models, which materially improved accuracy on higher-value, complex tasks and helped usher in the era of AI agents. This shift from simple prompt-response (next-token prediction at maximum speed) to deliberate reasoning was enabled by post-training reinforcement learning (RL), which teaches models to plan, use tools and evaluate intermediate steps against a goal.

In practice, RL provides the model with an environment and action space, tools it can use within that environment, and a reward signal aligned to desired outcomes. For example, in a coding environment, the model might have access to a code interpreter, the ability to write and execute code and rewards based on whether the program runs and produces correct results. This paradigm shift spurred major model providers to invest heavily in post-training RL and ultimately paved the way for early agentic applications that we are familiar with today, which includes deep-research, developer/coding agents and computer-use agents.

While frontier labs pioneered this work, an emerging ecosystem has formed around providing enterprises with reinforcement learning capabilities to train custom agents for real-world tasks. The key enabler is high-fidelity RL environments, which provide simulated workspaces with realistic observation and action spaces, integrated tool access (e.g., code interpreters, web search) and scalable infrastructure for iterative learning. These environments can be tailored to mirror almost any knowledge-work task, and early simulations span workflows from Excel spreadsheets to Salesforce dashboards. The ecosystem is converging on reinforcement learning as a service (RLaaS): managed platforms that abstract the infrastructure complexity for developing, training and deploying agents.

While there is industry enthusiasm around the potential for RL environments to solve the "last-mile" challenge of agent accuracy for domain-specific use cases, a hurdle to widespread adoption revolves around creating reliable evaluations or "rewards" (e.g., feedback signal that tells the agent how good or bad its last action was). Consequently, early RL successes are concentrated in domains where rewards are easily verifiable, such as code and mathematics (e.g., does the code run). For more nuanced, subjective tasks (e.g., generating investment memos or legal briefs), current methodologies pair SME-defined natural language rubrics with "LLMs-as-a-judge" evaluators that score agent actions based on the human provided rubric. An emerging industry view is that differentiated value is migrating from base models themselves to reward design, precise evaluation, and high-fidelity RL environments. As that shift takes hold, RL is moving beyond major labs, enabling enterprises to train purpose-built agents for domain-specific workflows.

Market and industry perspectives

All major model providers (OpenAI, Anthropic, Google, xAI) are investing heavily in RL to both improve reasoning in their general model capabilities and target high-value vertical domains. This investment can be seen through Anthropic's plan to spend more than $1 billion on RL environments over the next year to train models in complex professional workflows.[4]

Additionally, we are seeing the launch of managed reinforcement tuning services that allow developers to leverage RL for vertical-specific tasks, moving beyond generic model usage.

Traditional data labeling companies, which have historically supplied major AI companies with custom datasets for model training, have expanded their product suites to offer RL environments for both AI model providers and enterprises. Their differentiation strategy involves leveraging specialized human expertise to construct these environments and design appropriate reward models tailored for verticalized, industry-specific tasks.

Along with the established players, a specialized ecosystem of well capitalized startups has emerged to provide Reinforcement Learning as a Service (RLaaS) for enterprise-specific workflows, independent of any single model.

Consolidation is already occurring in this space. Leading GPU cloud providers are acquiring specialized RLaaS startups to expand their offerings beyond large model training and inference. These acquisitions bring targeted tools for developers to build and deploy agentic workflows directly on cloud platforms, helping providers differentiate their services and broaden their customer base beyond major AI labs. Similarly, AI inference platforms are acquiring companies focused on post-training and customization, allowing them to move beyond efficient, low-latency model serving and strategically position themselves for deeper model optimization and broader user capabilities.

[4] Zeff, M. (2025, September 21). Silicon Valley bets big on "environments" to train AI agents. TechCrunch. https://techcrunch.com/2025/09/21/silicon-valley-bets-big-on-environments-to-train-ai-agents/?secureweb=ONENOTE#:~:text=in%20RL%20environments%20to%20keep,environments%20over%20the%20next%20year.

Context engineering for the end-to-end software development lifecycle

Context Engineering will shift the way software developers interact with AI coding tools.

The common context engineering techniques of skills, knowledge graphs, episodic memory and prompt engineering are being applied throughout the entire development lifecycle, ultimately impacting the quality of code developed. Developers can build skills on how an agent should perform a specific task, especially repetitive tasks that would be standard across any deployment (e.g., using a specific file or language, applying internal company frameworks for code quality, leveraging specific API clients, or auto deploying procedures). Skills can dramatically cut down the time a developer spends doing these specific tasks and provide consistency across an organization.

While knowledge graphs have always played a role in understanding the connections between sources of data, the sophistication of knowledge graphs has evolved dramatically over the past year. AI developer tools are building knowledge graphs that generate detailed wiki pages with data flow and architecture diagrams of code repositories to help developers understand what effects code changes will have. These wikis are auto-updated based on human edits but can also be manually edited to be more accurate based on additional context given by the developer. Additionally, most AI dev tools are stateless by default – they do not retain information from previous sessions or prompts, causing friction for developers as they interact with a model since it loses previous context on an intended outcome. As a result, we are starting to see more tools enable episodic memory that remembers coding session history and stores completed conversations as searchable archives.

Skills, knowledge graphs and episodic memory can be triggered through effective prompt engineering. While prompt engineering is not a new practice, knowing how to use it in combination with the other context engineering techniques makes it much more powerful than it has been. Developers need to know when to call on a specific skill or even protocol (e.g., MCP) and use the right language to activate the model to recall a memory or identify a connection in a knowledge graph. Used in combination, context engineering will fundamentally change the speed and efficiency at which developers build and deploy software. However, individuals and enterprises alike need to understand that leveraging these techniques, despite its focus on increasing efficiency, may increase the overall cost of development as agents are increasingly used throughout the process.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

Leading players have launched AI development tools for deeper understanding of codebases.

Some code review solutions are embedding more code understanding and search capabilities into their solution to provide developers with more context into their code for more efficient code review. Other end-to-end software development platforms are building software delivery knowledge graphs that will provide additional context to AI agents throughout the entire SDLC. Elsewhere within the market, there is focus on offering a strong context engine to support better code generation.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Evolution of AI protocols

AI protocols enable standardized, cross-platform communication and collaboration between autonomous agents and tools, streamlining integration and scalability.

As AI shifts from isolated models to interconnected autonomous agents and multi-agent systems, AI protocols have emerged to provide standard ways for agents to communicate, exchange context, access tools/data, and collaborate across platforms. Announced in late 2024, Model Context Protocol (MCP) has emerged as the leading standard – but new alternatives and complementary protocols are proliferating, creating a rich and evolving ecosystem. Their proliferation reflects the increasing need for interoperability, portability, and governance as agentic AI moves into production, and suggests that AI protocols will continue to evolve over time.

In practice, this means developers no longer need to build custom, one-off integrations for every model, tool or agent. Instead, AI systems can plug into shared protocols that enable cross-platform tool access, shared memory and context, standardized communication and modular workflows – making agentic AI more portable, scalable, and maintainable.

In the past year, the industry has seen numerous protocols released from leading AI providers with slightly different approaches. While MCP is focused on connecting models to tools, protocols like Agent2Agent (A2A) and Agent Communication Protocol (ACP) define how agents communicate, share context and coordinate tasks. Agent Payments Protocol (AP2) provides a common language for secure transactions between agents and merchants; and there are even more protocols intended to make multi-agent deployment, observability, and governance feasible across enterprises. Given the explosion of AI protocol adoption, notably MCP, vendor solutions are rapidly building out MCP servers for users to connect to and many vendors today have registries and toolkits to simplify the installation and setup of MCP tools.

While these protocols provide a level of standardization, more tools connected to an agent or AI system may consequently cause inefficiencies. For instance, MCP passes all information through AI workflows no matter how simple or complex a task is, including information like tool definitions, descriptions, etc. This increases the token consumption and costs that may ultimately slow down an agent's performance. There are emerging techniques that suggest MCP servers as code APIs may be more efficient, but the industry continues to rapidly evolve and expect this space to mature over time.

Market and industry perspectives

In December 2025, The Linux Foundation launched a new initiative, the Agentic AI Foundation (AAIF), co-founded by Anthropic, OpenAI and Block – with broad backing from leading tech players including Google, Microsoft, Amazon Web Services (AWS), Cloudflare and Bloomberg. The founding projects donated to AAIF include MCP, OpenAI's project-instruction format AGENTS.md, and Block's agent framework goose – signaling a collective move toward unified, open, vendor-neutral agent standards.

According to Gartner's 2025 Software Engineering Survey, by 2026, 75% of API gateway vendors and 50% of iPaaS vendors, will have MCP features.[5]

In June 2025, Google Cloud donated its Agent2Agent protocol to the Linux Foundation and formed the Agent2Agent project with AWS, Cisco, Microsoft, Salesforce, SAP, and ServiceNow to collaborate and foster an open and interoperable ecosystem for AI agents with the A2A protocol.

[5] Zarecki, I. (2025, November 23). MCP Gartner Insights for 2025. Delivering Data Products in a Data Fabric & Data Mesh. https://www.k2view.com/blog/mcp-gartner/#:~:-text=By%202026%2C%2075%25%20of%20API,stability%20and%20address%20new%20requirements.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Agentic site reliability engineering

Autonomously monitor, diagnose, and help remediate system issues, enhancing traditional practices with pattern recognition, root cause analysis and integrated observabilit.

Agentic Site Reliability Engineering (SRE) represents a transformative shift in the way enterprises approach system reliability and observability. Leveraging the power of large language models (LLMs) and AI-driven agents, Agentic SRE tools are designed to autonomously monitor, diagnose, and remediate system issues, significantly enhancing the efficiency and effectiveness of traditional SRE practices.

At the core of Agentic SRE is the concept of autonomous agents. Capable of operating 24/7 to plan and execute actions on behalf of users, these agents utilize advanced capabilities such as pattern recognition, anomaly detection, and root cause analysis (RCA) to provide detailed insights into system performance. For instance, they can autonomously search through logs and databases, similar to the investigative process a human SRE would undertake, to identify and diagnose issues.

While the current scope of auto-remediation capabilities is not fully autonomous – requiring human initiation for executing recommended fixes – Agentic SRE tools offer significant advancements in RCA. They provide comprehensive knowledge graphs, confidence ratings, and documentation to support their hypotheses, thereby streamlining the troubleshooting process. Additionally, these tools integrate seamlessly with various observability solutions, code repositories, and IT service management platforms, allowing for a holistic view of the system's health and performance.

Despite their longer-term promise, agentic SRE tools are still evolving. For instance, they currently lack out-of-the-box capabilities to measure essential SRE metrics such as burn rates and error budgets. While the goal is to excel in reducing Mean Time to Resolve / Repair (MTTR), a critical metric for assessing system reliability, they are only part of the way there, but the industry expects their capabilities to mature over time.

Market and industry perspectives

The market for agentic SRE solutions is rapidly expanding as enterprises seek to enhance their system reliability and reduce downtime. The global market for AI-driven observability and reliability tools is projected to grow significantly, driven by the increasing complexity of IT environments and the need for more efficient incident management.

Key players in the agentic SRE space include vendors like Deductive, Traversal, Resolve.ai, and Dynatrace. Each of these companies is investing in unique capabilities to differentiate their offerings. For example, Deductive focuses on code-aware observability, leveraging source code analysis for RCA, while Traversal emphasizes in-depth RCA with confidence levels. Resolve.ai is exploring the integration of tribal knowledge into their agents, mimicking the expertise of human SREs. Dynatrace, with its Davis AI Copilot, is currently being implemented at JPMC, although its capabilities are limited to data within the Dynatrace platform.

Investments in agentic SRE technologies are focused on enhancing auto-remediation capabilities and expanding integration with various data sources. Companies are also exploring ways to incorporate historical incident data and tribal knowledge to improve RCA accuracy and effectiveness. The ability to integrate with any observability vendor and build comprehensive knowledge graphs is a key differentiator for these tools, enabling them to provide a more complete understanding of system health.

As the industry continues to evolve, agentic SRE tools are poised to play a crucial role in modernizing system reliability practices, offering enterprises the opportunity to reduce downtime, improve system performance and ultimately enhance customer satisfaction.

Observing AI

Traditional observability (metrics, logs, traces) provides a foundation, but new metrics and frameworks are needed to measure AI-specific behaviors, model drift, and compliance.

As the industry continues to adopt AI and leverage more agents for workloads, observing LLMs and agentic workflows remains paramount. It is critical to understand what these agents are doing, how they are performing, and the impact they may have to surrounding systems and applications.

Traditional observability pillars serve as a starting point for understanding AI behavior. Metrics, events, logs, and traces have become the defacto way of measuring health and behavior for applications and infrastructure through OpenTelemetry (OTel) standards. Distributed tracing (traces), which captures the flow of a request as it moves through parts of a system, is not as widely adopted as logs or metrics, but may ultimately be the best method to observe AI workflows. Traces can track what an agent did, what inputs/outputs were processed, tools or pre-set instructions it called on, how decisions flowed from one to another, and provide verification that the agent accomplished each step it planned to do.

However, new metrics and frameworks are needed for deeper insights. AI workloads introduce a new level of measurement we have not had before. LLM performance can be measured by number of tokens used for a specific task, response quality (accuracy and completeness), model drift and data quality checks, and the safety and compliance of the AI interaction. These metrics are starting to be offered by an emerging wave of startups focused on AI observability, but also industry incumbents as well. Ultimately, the most effective measurement frameworks will likely integrate real-time signals and alerts from emerging players with long-term performance tracking of agents provided by traditional vendors.

With metrics on how AI models are performing in production, companies are also looking at ways to gather metrics pre-production through model evaluation. Model evaluation involves testing models and understanding how its performance may impact workloads in production, improving the speed at which models are deployed and improving the risk of models failing. Observability metrics for both pre-production and post-production are necessary to understand the end-to-end spectrum of model behavior.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

Incumbent observability vendors are releasing LLM Observability into their platforms to develop, evaluate and monitor LLM apps.

The OpenInference specification emerged this year, which is a set of conventions and plugins complimentary to OpenTelemetry that enables tracing of AI applications. It is designed to provide insight into LLM calls and the surrounding app context (e.g., vector DBs).

AI Observability vendors have seen significant fundraising rounds in the past few years; and a number of recent acquisitions have highlighted the need to embed LLM observability into existing products.

Data-centric security policy

Securing and protecting data wherever it moves will be foundational to AI and agentic systems.

AI-driven data flows pose new challenges for governance and security. As organizations inject data context into generative AI and multi-agent systems, data is no longer static – it flows freely and continually transforms through model prompt and response, ephemeral memory, vector embeddings, knowledge graphs, and agentic operations. AI systems turn enterprise data into derived representations of source data. Data synthesized by an agent from multiple sources can also generate an output of higher sensitivity than the source data. Agent memory and knowledge graphs can bring valuable context but can also store or infer data – for example sensitive implicit relationships – which can be unintentionally exposed.

Protecting data now means securing its movement, not just its storage location. Given this shift, approaches to protecting data are evolving from securing data where it exists to securing how it moves. Data-centric security is an architectural shift that treats data, access to data, and the flow of data as the primary control plane. Data centric security models embed security policy, governance and telemetry directly into data objects so that classification, access permissions, lineage and data protection “travel” with the data wherever it goes.

Data controls must adapt dynamically to support granular context and intent of interactions. Attributes defined at the data object level with sensitivity labels and usage constraints could enable any agent, model or system interacting with data to comply with those rules at retrieval time and also persist those attributes downstream as new data is created. Sources of context like embeddings and graph edges will require the equivalent of table and row level security to enforce dynamic policies – for example, “agents of type X cannot act on data of type Y”. There is also an emerging focus on securing derived data (e.g., embeddings, memory) by scanning these stores for sensitive or potentially maliciously poisoned data.

Data Loss Prevention (DLP) models are also evolving to support data risk detections based on end-to-end lineage – tracking data from where it originates, where it flows and how it transforms – in addition to where it may egress the organization.

Finally, advanced techniques can protect data in agentic systems. There is increasing use of confidential computing and runtime isolation to create trusted boundaries for model inference and agent execution to ensure that sensitive raw data and context never exists unencrypted in memory outside a secure enclave.

Data Security Posture Management (DSPM) tools are expanding to secure data in transit. Data security platforms, commonly referred to as DSPM tools are extending from scanning and classifying traditional data stores at rest to scanning data at source, AI model inputs / outputs, vector embeddings, agent memory and data in motion to identify existing data risks and prevent unwanted sensitive data from entering into AI systems from the start.

These DSPM players and new emerging startups are also building capabilities to observe data movement and lineage throughout its lifecycle to better identify and triage potential risks (e.g., privacy impacts, data leakage).

Cloud runtime isolation solutions and confidential computing offerings are providing dedicated session isolation to ensure agent state, tool operations and credential access remain completely compartmentalized. When a session ends, the entire environment can be terminated and memory sanitized, minimizing the risk of data persistence or cross-tenant contamination.

Companies have also introduced new models to protect the consumption of proprietary data. Dynamic tokenization capabilities are also emerging to support direct model interactions while safeguarding sensitive data from model providers.

Agentic identity access management

Redesigning traditional human-centric IAM to allow dynamic, intent-aware and auditable authorization for autonomous AI agents.

With the rise of autonomous AI agents, traditional, human-centric identity and access management (IAM) is evolving. As agents increasingly perform actions on behalf of human users, concepts like Know Your Agent (KYA) are quickly gaining traction in customer-facing use cases, ensuring autonomous actions are legitimate, secure, and acting within authorized bounds for financial and sensitive transactions, and preventing fraud by confirming the AI's origin, permissions, and owner.

For the workforce, assigning an identity or authenticating an agent is just the beginning. When an agent is performing actions on behalf of a human employee, identity tokens must explicitly bind together three elements: the agent's identity, the identity of the original requester (human, software or agent) and the intent or context of the request (e.g., which resource is being accessed and why).

To prevent agents from overstepping their bounds, access can be downscoped, i.e. the agent only receives the subset of the human user's permissions that allow it the least privilege that is strictly required for the specific task, granted just-in-time and revoked immediately after use. This structure minimizes overtly permissive entitlements by design so that human approval can be reserved for only the most critical or sensitive agent actions. In multi-hop workflows, where agents invoke other agents or services to perform downstream actions, authorization must persist across hops to the very last step maintaining a clear trace of identity, origin and scope of the original request.

With the rise in agents in SaaS applications, IAM must evolve beyond traditional perimeter-based controls to support secure SaaS-to-SaaS interactions where trust, scope and intent can persist without a shared IAM layer. Enterprises must redesign architectures and ownership models to enforce business policy across internal and third-party agentic systems. In parallel, industry standards and industry-wide adoption of agent IAM best practices will be critical to driving and shaping safe, transparent and reliable agentic interactions.

Market and industry perspectives

Recent research by Salesforce predicts that AI Agent adoption is expected to jump 327% over the next two years.[6] With the rise in AI agents, cloud providers are formalizing the recognition that agent identities will become a distinct control plane for AI agents. For example, AWS and Microsoft have products to govern credentialing, policy and audit lifecycle for AI agents on their platforms.

Large security platforms have been actively consolidating identity capabilities.

Networking players are expanding into SaaS-to-SaaS and API-level access governance to create gateway-style cross-application controls that can help federate access for agents across platforms. This activity signals a shared recognition that identity will become the primary enforcement layer for agentic AI.

A number of emerging startups are innovating in the space to build control planes to manage, govern and remediate access for agents. While early, the space is evolving quickly to deliver solutions for scaled enterprise adoption.

[6] HR Leaders to Redeploy a Quarter of Their Workforce as Agentic AI Adoption Expected to Grow 327% by 2027. (2025, May 5). Salesforce. https://www.salesforce.com/news/stories/agentic-ai-impact-on-workforce-research/

Human risk management

Embedding human behavior as a security signal to limit exposure resulting from both human errors and susceptibility to adaptive, personalized threats.

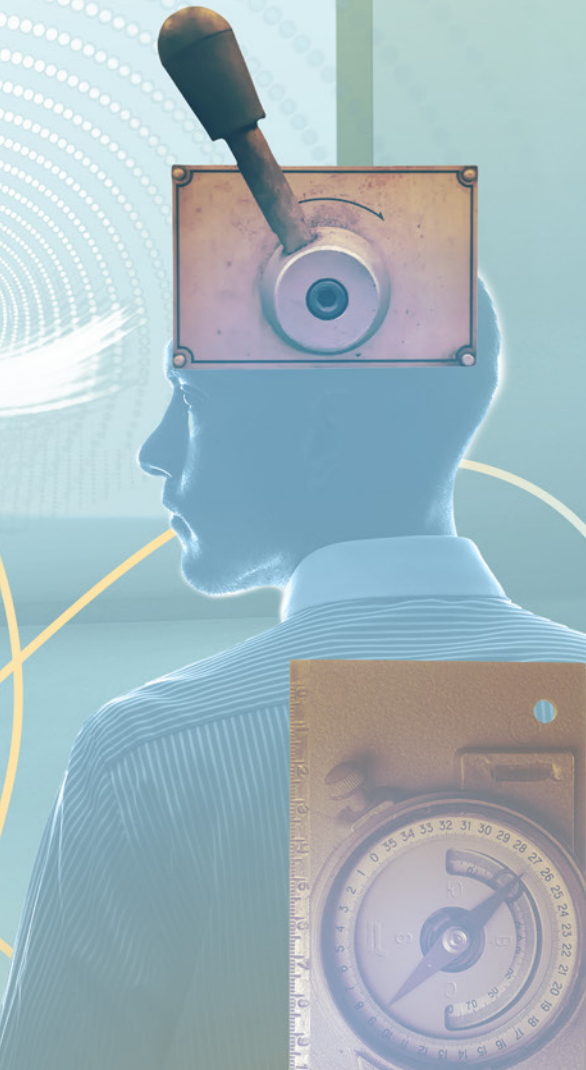
Humans remain one of the most consistent sources of cybersecurity risk, prone not only to malicious targeting like social engineering or fraud but even well-intentioned human errors like sharing a sensitive document with the wrong recipient or accessing unauthorized tools that can expose or leak sensitive data and create compliance issues.

One area of industry focus is minimizing susceptibility to social engineering threats through AI-powered security training, i.e., platforms that enable enterprises to simulate highly realistic phishing campaigns using deepfakes, cloned voices and context-aware prompts to mimic modern multi-modal attacks to augment awareness among the workforce.

Yet another avenue that AI has opened up for attackers is SEO poisoning, i.e., manipulating search results so a company's employees on the internet are redirected to malicious websites, fake login pages or corrupted tools during routine work. These tactics are particularly effective as they exploit legitimate user intent rather than deception alone. Defending against such tactics combines behavioral signals with browser telemetry and business context to warn users or block access in real-time when high-risk patterns are observed.

Many solutions are also emerging to manage fraud risk across users, customers and clients such as detecting fraud patterns in day-to-day communication channels or flagging abnormal behaviors in customer support flows. In all cases, human behavior becomes a security signal in and of itself.

As the threat landscape evolves, so too do the tools that help manage human behavior as part of the security architecture. Enterprises are looking to embed behavioral context directly into real-time policy enforcement decisions. This goes beyond role-based access and policy checklists. For example, rather than sweepingly banning any and all attachments to external email addresses, personal documents being shared with the employee's own personal email could be selectively allowed. Conversely, deviations from known behavioral baselines should trigger escalations or controls, even if policy would otherwise allow the action. This approach underpins dynamic policy-enforcement blending human behavior with telemetry from browsers, file systems and identity platforms.



This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

In a recent IBM survey, 74% of Chief Information Security Officers (CISOs) ranked human error as the top risk to their organizations' cybersecurity.[7]

Large platform providers are recognizing that human-targeted attacks require ecosystem-level defenses. Google's recent partnership with Doppel, an emergent brand protection solution, for example, focuses on combining telemetry for faster identification and takedown of risky domains. At the same time, a new wave of solutions are focused on using AI to outpace the complexity of AI-powered attacks by using AI-powered simulations to train the workforce against phishing attempts and fraud.

Beyond the workforce, multi-channel fraud prevention solutions are working toward applying similar behavioral and real-time signals to detect customer-facing fraud across payments, messaging and digital media.

This shift is evident across various layers of security. Startups are positioning products around understanding baseline human behavioral patterns and AI powered risk reasoning solutions to allow organizations to act on human behavior before it comes an incident. Further, companies are framing the internet browser as the place where human intent, navigation patterns, and interactions can be observed and used to inform policy enforcement decisions in real-time.

[7] Gregory, J. (2024, August 15). CISOs list human error as top cybersecurity risk. Ibm.com. https://www.ibm.com/think/insights/cisos-list-human-error-top-cybersecurity-risk

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making