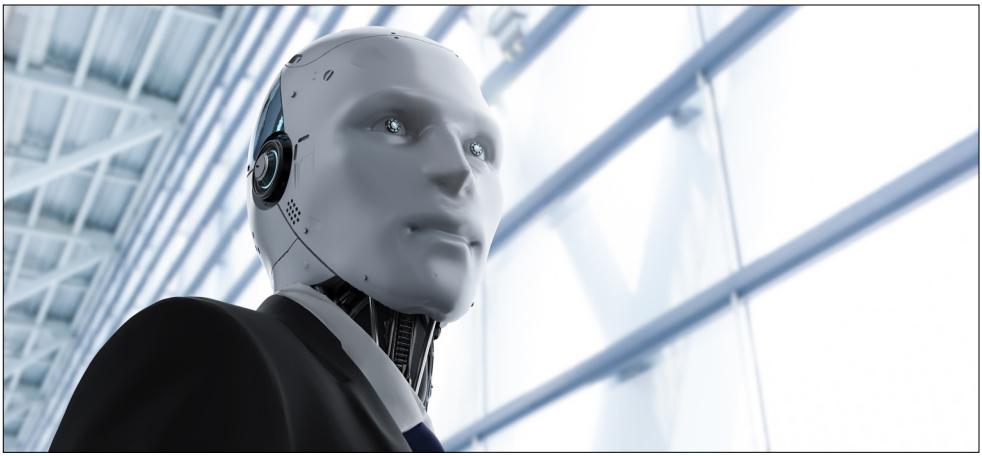
Enterprise Spotlight: The Promise and Perils of Agentic AI



Credit: Phonlamai Photo/Shutterstock

Generative AI captured the imagination in its ability to interact with people around language and content. The next stage is agentic AI, which gives artificial intelligence more autonomy to manage end-to-end processes without human involvement. That could provide huge efficiencies but also risk soaring costs and deep security threats not seen until it's too late. This issue helps IT navigate the emerging agentic AI world.

CONTENTS

From the editors of Foundry's enterprise IT sites:

CIO CSO COMPUTERWORLD InfoWorld NETWORKWORLD

How AI agents will transform the future of work

BY ISAAC SACOLICK, INFOWORLD.COM

AI agents are already reengineering software development, business processes, and customer experiences. Here's what you need to know.

At first, robotic process automation coupled with low-code platforms and orchestration tools propelled many organizations to increase productivity and scale business operations. Virtual agents and chatbots took automation one step further by enabling a conversational experience. Then, large language models (LLMs), vector databases, retrieval augmented generation (RAG), and other generative AI innovations enabled new ways to summarize content, generate code using copilots, and answer questions conversationally.

AI agents combine automation, conversational experiences, and process orchestration capabilities to lead us to the next phase of generative AI evolution and digital transformation. They provide developers, business users, and others with a role-based partner, proactively automating steps and acting as knowledgeable collaborators in getting work done. Integrating genAI technologies with role-based workflows is a key opportunity to deliver transformational generative AI business benefits beyond productivity improvements.

"AI agents are fundamental to practical, measurable applications of generative AI in the enterprise," says Simon Margolis, Associate CTO of AI/ML at SADA. "Their unique ability to act — that is, write data and make API calls — represents a huge advantage in how businesses cannot just gain information from their AI tools but use those tools to perform actions that are otherwise poor uses of human time."

Platforms such as Appian, Atlassian, Cisco Webex, Cloudera, Pega, Salesforce, SAP, ServiceNow, and Workday announced AI agent capabilities in 2024, while public cloud agents such as Amazon Q Developer are embedded in the developer experience. The AI agent market size was valued at $3.9 billion in 2023 and is expected to grow at an annual growth rate of 45.1% from 2024 to 2030.

One reason for the massive interest in AI agents is that they bring expertise and automation to the workflows users perform regularly.

Margolis adds, "Action agents are pivotal in helping organizations realize the measurable benefits of generative AI, whether it's a simple sales agent asking a salesperson basic customer information to automatically make entries into a CRM system or a medical agent providing information about a patient and updating their records post-visit."

RULE-BASED CHATBOTS VERSUS AI AGENTS

Virtual agents and chatbots are often rule-based approaches to help users solve a handful of basic problems. For example, IT services management (ITSM) chatbots often address common service requests such as password resets and unlocking accounts, but then redirect users to FAQs and knowledge bases for more complex requests. ITSM AI agents can perform more sophisticated tasks like predictive incident management, intelligent ticket routing, and problem root-cause analysis.

"AI agents are changing the game across industries by automating tasks, solving problems, and improving workflows," says Abhi Maheshwari, CEO of Aisera. "Unlike standard chatbots, these agents can reason, plan, and take action independently. They're used in areas like tech, manufacturing, legal, retail, education, and government."

Many platforms now have sidebars on web pages and other user experience elements where users can interact with AI agents around their work. Sometimes, the agent presents information proactively so people can take action. At other times, they lend expertise and share data-driven insights with the employee while performing their work.

"To the user, the app interface of chatting with a bot feels familiar, as if they are chatting with one of their colleagues or in a group chat," says Gaurav Kumar, VP at LatentView Analytics. "The difference is that the chatbots' responses are succinct, insightful, instantaneous, and derived from a vast corpus. An additional benefit of this powerful technology is the ability to control who gets to see what using role-based access controls."

Parul Mishra, VP of product management in digital labor at IBM, says enterprise AI agents and assistants will coexist along a continuum, with AI assistants executing prescriptive tasks while the more self-directed AI agents reason through complex problems and execute multi-step plans to solve them. Mishra says, "While the capabilities of AI agents across specific domains

will vary depending on industries and use cases, the key point is that those agents will work in collaboration with AI assistants alongside a suite of other tools to transform a chatbot experience into a multi-dimensional system that can plan, test, write, and autonomously implement solutions."

AI AGENTS REQUIRE HIGH-QUALITY DATA

To be helpful AI agents need accurate, relevant, and up-to-date information to provide accurate answers. Before leveraging AI agents, data leaders should learn what data the AI agent accesses and validate its quality.

"AI agents are autonomous systems and workflows that help automate complex tasks and decision-making; however, their efficacy relies on having access to high-fidelity data as input," says Abhas Ricky, chief strategy officer at Cloudera. "If agents take action on unverified data, it can inadvertently introduce errors or inefficiencies into critical operations, ranging from fraud detection to supply-chain management. Ensuring trusted data integrity and accessibility through robust data architectures allow agents to make accurate, impactful decisions."

Organizations should update their AI governance and data governance policies to include AI agent use cases. Now is also the time to review whether data pipelines need performance or other operational improvements. Larger enterprises looking to develop AI agents should review data fabrics to simplify access to data sets across SaaS, public cloud, and data centers.

"We'll also see an increase in the need for roles focused on AI governance, helping to ensure that virtual agents don't go rogue with unintended bias, which could result in serious implications for brands," says Don Schuerman, CTO of Pega. "Ultimately, the future of work with virtual agents provides more opportunities for enterprises to quickly help their customers while enabling their employees to be more productive with higher-value work, all while operating at increased levels of efficiency and effectiveness."

AI AGENTS CAN IMPROVE THE EMPLOYEE EXPERIENCE

AI agents offer productivity benefits, but more importantly, they can add joy to a user's workday. Instead of being bogged down by all the technical minutia needed to complete a task, employees can focus on what needs to be done while the AI agent collaborates on the implementation.

"AI agents will not only simplify human resource and finance processes, but they will transform how work gets done," says

David Somers, chief product officer of Workday. "By automating routine tasks such as submitting expenses, creating job descriptions, and scheduling, AI agents will allow employees to work more efficiently and focus on higher-value activities. This will help organizations save time and make more informed decisions without adding complexity to their workflows."

Some quicker wins for deploying AI agents involve taking the drudgery out of work, especially for infrequently performed tasks requiring data entry.

Using HR and finance AI agents can improve everyone's productivity because all employees submit forms and reports to these functions. Another quick win is in customer success and support, where providing accurate and timely answers can be challenging when there's a lot of information to review.

"A key use case is semi-autonomous customer success agents, which augment and bring value to customer success teams by automating tasks instead of just advising or assisting on them," says Raju Malhotra, chief product and technology officer at Certinia. "AI agents also play a big role for customer support, resource managers, and project managers by amplifying the impact they would not be able to have otherwise."

Technologists who want to understand the value of AI agents should review how they work in ITSM and DevOps practices. While copilots can help generate code, AI agents aim to perform tasks and reduce complexity across the full software development and deployment life cycles (SDLCs).

"Virtual agents are already transforming DevOps practices across all industries by providing productivity gains across all five phases of the SDLC," says David Brooks, EVP of evangelism at Copado. "We are in a transition period from copilots to true agents, where these virtual team members can not only provide advice but perform actions on your behalf. Agents will help free up people to focus on prototyping with the help of a plan and build agent until results meet expectations, and then DevOps agents can test, release, and operate the solution."

AI AGENTS FOR INDUSTRY-SPECIFIC USE CASES

Many currently available AI agents target departmental workflows and aim to improve productivity. AI agents aiming to solve industry-specific workflow challenges may prove to be of even greater business value as they can directly impact costs, quality, scalability, and customer experience. These AI agents tap into the enterprise's data and subject matter expertise to solve

real employee and customer challenges while improving experiences.

"Retail, manufacturing, and any industry with large product inventories will benefit from virtual agents that combine AI, an understanding of the buyer's needs, and deep context awareness for thousands if not millions of products," says Jonathan Taylor, CTO of Zoovu. "This solves two of the biggest issues in online buying today — choice overload and the high cost of returns due to low-quality product search and discovery."

AI agents that solve customer pain points and pinpoint relevant information from big data sources can provide business value, especially when AI agents are personalized to user roles and interests.

"In retail, virtual agents that understand a shopper's needs will interact with retailers' websites using machine-to-machine protocols to buy products based on the customer's interests, price, size, and other factors," adds Taylor.

The growing list of opportunities to use AI agents to accelerate industry 5.0 target work safety, predictive maintenance, and improving quality. Advancing in these and other areas will require IT to accelerate efforts to centralize and cleanse operational data.

"Manufacturing and physical industries will lag behind knowledge industries in developing and adopting AI-powered agents due to the lack and complexity of data infrastructure," says Artem Kroupenev, VP of strategy at Augury. "That being said, we're starting to see industrial AI agents take shape around machine health diagnostics and production health recommendations because these types of solutions are narrowly focused on specific high-value use cases and come as a full-stack package with built-in hardware and data infrastructure required to make these AI agents effective."

AI agents in manufacturing will add to existing machine learning capabilities and provide workers with two-way human-to-machine interfaces using natural language prompting.

"With AI agents for machine and production health, we're already seeing a preview of what's possible for AI in manufacturing: dramatic impact on productivity, increased capacity and sustainability, increased safety and augmentation of human decision-making and collaboration on the production floor," adds Kroupenev.

Kevin Miller, CTO of Americas at IFS, shares a common problem in manufacturing that frustrates workers and impacts operations and costs. "Factory workers and field technicians often spend significant time troubleshooting unique problems not documented in

any manual," says Miller. "AI-enabled customer support tools enable predictive and prescriptive maintenance to minimize unexpected machine downtime or service interruptions."

RISKS AND OPPORTUNITIES USING AI AGENTS

AI agents are relatively new, which brings both opportunities and risks for early adopters.

Eilon Reshef, cofounder and chief product officer at Gong, says AI agents are currently unsuitable for more comprehensive tasks that require extensive knowledge and history. Reshef says, "Asking an agent to share key takeaways from a call is not the same as asking it to plan next quarter's revenue strategy, and the potential for unpredictable outcomes grows with the amount of autonomy you give the agent. Today's AI agents can be considered revolutionary for internal process automation, but enterprise leaders may need to think twice about having them take action that can be seen or felt outside of the organization."

If AI agent capabilities improve at the same velocity as LLMs, we will see them become more autonomous and capable of agent-to-agent integrations.

Maheshwari of Aisera says, "The next big leap will be autonomous AI agents that can manage entire processes without human intervention, leading to major productivity boosts. These agents will handle routine tasks, allowing organizations to work smarter and faster. While challenges remain, this technology is set to transform how we work."

"Eventually, AI agents will interact with each other behind the scenes, automating many human processes as one agentic system," adds Deon Nicholas, CEO of Forethought. "We'll likely see a network of AI agents, similar to how the Internet today is a network of computers. Each person will have an AI agent that can book haircuts or order groceries, and those shops will have customer-experience AI agents to receive the orders, process returns, and take actions."

Platforms are already delivering autonomous and AI agent integrations. One example is the recently announced Workday and Salesforce integration between their AI agents.

Our mainstream view of workflow focuses on task automation, process orchestration, machine learning intelligence, and human-in-the-loop decision-making. Adding AI agents to this mix will usher in a new era of digital transformation by reengineering business processes and customer experiences. ■

How will AI agents be priced? CIOs need to pay attention

BY GRANT GROSS, CIO.COM

Agentic AI, the more focused alternative to general-purpose generative AI, is gaining momentum in the enterprise, with Forrester having named it a top emerging technology for 2025. Since then, several organizations have begun using the technology, and major vendors such as Salesforce and ServiceNow have offered AI agents to customers.

Agentic AI focuses on performing specific tasks and emphasizes operational decision-making instead of the content generation often associated with gen AI tools.

The technology is in its early days, and several questions remain open — chief among them, how AI agents will be priced. So far, no agreement exists on how pricing models will ultimately shake out, but CIOs need to be aware that certain pricing models will be better suited to their specific use cases.

Salesforce, for example, offers three pricing models: one that includes 1,000 Agentforce "conversations" free with its Salesforce Foundations CRM service; another included with its standard success plan; and $2 per conversation a la carte. Salesforce defines a conversation as a customer sending at least one message or selecting at least one menu option or choice other than "end chat" within a 24-hour period.

The $2-per-conversation approach can include many back-and-forth interactions between a customer and Agentforce, says Ryan Schellack, senior director of AI product marketing at Salesforce. The company is focused on use-based pricing, with only one customer seat required to administer it, he adds.

LOTS OF PRICING MODELS TO CONSIDER

The per-conversation model is just one of several pricing ideas. In a recent LinkedIn post, Box CEO Aaron Levie outlines four agentic AI pricing models that could emerge.

First, vendors could base the price of AI agent tasks on the traditional work they replace, with a discount on the traditional labor price. "An AI agent performs a certain amount of work, and you pay for amount of time or units it took to do that

work," he writes. "Generally, it's a fair trade for the customer and provider."

Second, agents could be priced based on outcomes, with the price focused on the completion of a task. "This model allows for a simple relationship between what the customer needs and what they're paying to get accomplished," Levie writes. "It also has the benefit that as underlying AI costs drop over time service providers can extract more margin for this work."

A third way that AI agents could be priced is by calculating the underlying costs and charging a small markup, he says. "This can be great for technically savvy customers but has the risk of not being sufficiently abstracted from AI costs to hold value over time," he says. "Potentially good for customers, but maybe not for shareholder returns."

Finally, agentic AI vendors could offer a per-seat SaaS subscription model that gives users unlimited access to the agent, Levie says. "This model could be quite disruptive," he writes. "In areas where there are lots of seats used by users, it's possibly very strategic; in areas where there's only a small number of seats, you're likely giving up too much value."

More pricing models are likely to come forward, Levie says in an interview with CIO.com. These are "fairly exciting times to watch new business models in software emerge after a decade plus of limited changes," he writes.

CONVERSATIONS AND SUBSCRIPTIONS

A per-conversation model seems to be an emerging approach, says Sesh Iyer, managing director, senior partner, and North America regional chair at BCG X, Boston Consulting Group's IT building and designing group. Vendors could also charge a small price per audio input or output. Alternatively, a token-based consumption approach would bill tokens used for assistant API tools at the chosen language model's per-token input and output rates, he adds.

An early trend seems to be the SaaS model, with a per-conversation model emerging for infrequent users, says Ritu Jyoti, general manager and group vice president for AI, automation, data, and analytics research at IDC.

Outcome-based pricing could be tricky, she says, when it's still difficult to define a successful outcome in an AI agent intervention. Outcome-based pricing may lead to disputes between vendors and users over whether the desired effect was achieved, Jyoti says, although pricing based on resolutions can work well in customer-service situations.

"It is all dependent upon the features and usage volume," she adds. "What is

really desired from enterprises, as they kind of get into this whole adoption, is that they are looking for subscription-based pricing with tiered plans based on features and usage volume. The reason is because enterprises look for some predictability."

However, some experts see other price models emerging. Outcome- and cost-based pricing, with variations for specific use cases, are likely to catch on, argues Rogers Jeffrey Leo John, cofounder and CTO of DataChat, a no-code, gen AI platform for instant analytics. In comparison, current large language model pricing is a form of outcome-based pricing, with users paying for tokens processed or generated, he notes.

"For business users, outcome-based pricing is often the most intuitive," says Leo John. "This model directly ties the cost to specific outcomes or successful completions, making it easier to relate to the value delivered."

Cost-based pricing will also be appealing because it's straightforward to calculate, he adds. "By pricing based on the underlying costs of compute, latency, and throughput, this model provides clarity on how charges are determined and allows for more precise budgeting," Leo John says. "While it may lack the direct ROI alignment of the outcome-based model, it simplifies the financial planning process for users who understand and manage technical resources."

DANGERS OF CONSUMING TOO MUCH

While several pricing models may emerge, CIOs and IT leaders should beware of consumption pricing, says Jeremy Burton, CEO of Observe, an AI-powered observability platform.

"It all sounds good, but the challenge is that people get annual budgets and cannot tolerate variability," he says. "Everyone assumes if they move from subscription to some form of consumption, then they'll save money. That is until they see a spike and burn through half of their budget in a few weeks."

Some big vendors in the IT industry can demand consumption pricing, he says, "but from what I've seen at the app level it's a nightmare."

FOCUS ON YOUR BUSINESS NEEDS

Pricing will evolve as the agentic AI model does, and CIOs should explore the options that best fit their needs and their use patterns, experts say.

"AI costs have been dropping significantly, with the cost per unit of AI work decreasing and capabilities improving rapidly," Leo John says.

"Vendors may move towards hybrid models that combine cost-based transparency with performance-driven incentives. The ongoing advancements in AI will drive continuous evolution in how AI services are priced to remain competitive and aligned with market demands."

CIOs should consider specific use cases and desired outcomes with AI agents, Leo John adds. This assessment can determine whether an outcome-based pricing model or a cost-based model, based on operational expenses such as computing and throughput, is more suitable.

CIOs should also consider total cost of ownership, he says. "With new and improved models emerging almost daily, leaders must also account for the costs associated with retraining or customizing AI models in this rapidly evolving landscape," he adds.

CIOs and IT leaders must determine how their organizations will use the AI agents to determine the best pricing for them, adds BCG X's Iyer.

"They should assess what is available today with an in-depth understanding of pricing and volume, build forecasts for at-scale usage, and build scenarios for increase in unit costs —like cloud — with optionality to switch agents to prevent lock-in," he says.

Transparency and predictability will be the drivers for agentic AI pricing, says Box's Levie. The vendors that provide predictable prices — and outcomes — will win in the market, he predicts.

"You need a high degree of visibility into what you are actually going to be paying for, so you don't have these big, opaque systems when you can't really anticipate what's going to happen," he says. "You generally can't have a situation where you see a 10x spend increase, because something happened within the system that was a surprise." ■

The rise of agentic AI and its implications for security

BY STEPHEN KAUFMAN, CSOONLINE.COM

Agentic AI is on the rise, but so are its security risks. Here we reveal how to harness its transformative power while mitigating potential threats and navigating the complex security landscape its use presents.

The emergence of generative artificial intelligence (genAI) large language models (LLMs) — such as ChatGPT — has created an earthquake of change that has rippled through every industry and every business. We have all felt the shocks. But these shocks have introduced new capabilities, efficiencies, and possibilities. They have also shaken the existing structures, processes, governance, and operational activities to the core.

Most of us have been researching, using and incorporating genAI in some fashion in our organizations. But like all technology advancements, things move fast. Tools like ChatGPT and Microsoft Copilot are now almost commonplace. Many organizations continue to incorporate genAI into their applications while adding features and technologies that make the solutions more accurate and capable, such as model grounding or implementing retrieval augmented generation (RAG) patterns. Just as we are gaining an understanding of these capabilities, there is a growing trend that is obtaining more prominence: namely, agentic AI.

Agentic AI is a new trend of using AI in an iterative workflow approach that contains agents that act autonomously to achieve specific goals. They can make decisions and act without the need for human intervention or a human in the loop. They are always online, listening, reacting, and analyzing domain-specific data in real-time, making decisions and acting on them.

Gartner predicts that by 2028, one-third of human interactions with genAI will evolve from user-prompting LLMs to use interfacing directly with autonomous, intent-driven agents. This is a major jump ahead of the reactive AI assistants many users are now familiar with.

While this mix of autonomy and automation provides more advantages for certain applications, especially those used in dynamic and complex environments, the need for security is even more crucial now.

Like any new tool introduced into a system, it can also introduce new vulnerabilities. If, for instance, an agentic AI system is compromised, the decisions it makes autonomously could range from troublesome to disastrous and could include downstream effects. It's therefore important to examine the security considerations and areas that require extra security focus. You'll also want to investigate tools and patterns that can be utilized to make the system more secure.

Though you may not yet be aware of agentic AI, your development teams are likely already working with it. The time to understand agentic AI and implement security measures is now.

A BRIEF OVERVIEW OF AGENTIC AI

Let's dive into agentic AI a bit deeper to understand and set the baseline of the security measures that will be needed.

Agents, and agentic AI, are not the same as existing genAI services such as copilots or chatbots. Traditional genAI services accept a specific command or prompt and return a response. By contrast, agents work as part of a workflow process that acts based on the results returned from the agent.

Agentic AI brings together a set of tools, frameworks, and patterns to automate end-to-end business process workflows that enable AI and people to work together. Using agents, solutions can be built to deliver end-to-end results that can be pieced together to autonomously achieve the business objective.

AI agents are the foundational units of work in this architecture that drive each automation task. Each agent is designed to perform a specific, unique, autonomous task and seamlessly integrate back into the broader workflow. There it will deliver information from LLMs, internal business systems, data sources, or other AI services as well as data from external systems. These agents can interact with LLMs, allowing the creation of a variety of content. This includes generated code that can be executed by other agents further downstream in the workflow process.

The workflow process that directs these agents is the controller, which calls the agents required based on the unique sequence of activities established from the rules and decisions made with the data returned by each agent. In this process, the workflows can select the right agents to interact with appropriate APIs, determine the right sequence, and execute the processes to fulfill the business requirements.

As the organization builds out solutions and creates agents, it draws upon the strength of agentic AI, which is the ability to integrate external agents

that were not originally built on the platform or by the team. This not only allows teams to collaborate but also lets companies innovate and incorporate new technologies and capabilities without disrupting the existing solution.

Because of the autonomy and dynamic nature of the workflows and agents, there may or may not be human interaction in the end-to-end process (aka a human in the loop). However, there must always be the ability to control the process and operations (aka a human on the loop). We must be able to monitor the process, log what each agent is doing, log the data that each agent receives and returns from the workflow, and have the ability to shut the process down or override the operation.

You may be reading this and thinking that agentic AI is very similar to microservices. You are partially correct. Agentic AI is different in that you are placing the LLM in the control flow and letting it dynamically decide which actions to take. This makes agentic AI leaps and bounds more powerful and dynamic than microservices. So, if you have already put in place security and governance measures for your microservices solutions, you can start there and expand. However, what has been created so far is not sufficient to cover what is required for AI — and especially agentic AI.

Deciding how agentic AI will be integrated into your organization will require you to think about the complexities and touchpoints that must be covered. You will need to consider all the activities that need to be governed and monitored so teams don't create a black box solution. You will need to ensure there is no automation happening without direct human oversight and control. In addition, tools today are available that provide the ability for agents to be created and managed by nondevelopers, using low-code or no-code frameworks. This further drives our need to implement governance, security guidelines, and expansive testing to mitigate risks.

THE RISKS OF AGENTIC AI

There are risks associated with agentic AI. By understanding the risks, we can begin to put a strategy together for mitigation.

Unexpected behavior or problematic behavior

AI systems are non-deterministic and behave unpredictably or even, at times, counterintuitively. Agentic AI autonomy increases this risk. Agents may carry out tasks in ways that weren't anticipated. The decision-making and activity trail needs to be logged and transparent, otherwise it will be difficult for people to understand and harder to control or reverse the behavior.

Ethical concerns and dilemmas

Safety and ethical concerns must be at the forefront of the risk discussion. The autonomy of agentic AI agents raises the questions of potential misuse as well as unintended consequences. These questions need to be addressed to ensure we maintain trust and ensure the responsible use of these agents.

Bias is also essential to be aware of and protect against. When agentic AI systems make decisions, we need to be able to detect if there are biases that would lead to unfair or discriminatory results. Responsible AI standards are crucial not only for ethical concerns but also for transparency, explainability, responsibility, and visibility. Businesses and users won't move forward if they can't trust the solution.

Lack of human controls

As noted earlier, there must always be the ability to have control over the process and operations (human on the loop). People may think that introducing human controls will slow down performance and use that as an argument. We need to ensure that systems are being designed and implemented with the ability to monitor the process, log what agents are doing, what each agent receives and returns from the workflow, and shut the process down or override the operation.

We also need to review operational metrics so the system is performing within the goals and governance standards as well as remaining aligned with organizational goals.

Security risks

This is listed last because it is so important and a catch-all to the risks that arise. If an agentic AI system gets hacked, serious consequences can arise. First, we need to consider the time it takes to realize a hack has occurred and identify that it is a hack. Even a small unintended change or manipulation can have large consequences. Second, because of the autonomous nature of the system, new security vulnerabilities are introduced that need to be addressed. Last, the largest security risk is putting a system into production without monitoring, logging and controls. These are not bolt-on-afterward activities.

We need to address specific challenges such as testing for hallucinations, prompt injection attacks, and unfiltered user-provided text directly into prompts.

These risks are important to understand and mitigate. It is important to be leading the activities around controls and governance. According to security firm Portal26, 58% of organizations are concerned about the lack of visibility

into the unsanctioned use of genAI. Organizations need to lay the foundation regarding the activities and usage, but even more so, they need to be vocal about the rules of use and processes to use genAI and agentic AI in production. Without a process in place, and an understanding of where genAI and agentic AI are used, leaders will be wary of the liability. ISMG's Generative AI Study showed that 55% of leaders lack an understanding of how AI is and will be regulated and are seeking guidance.

It is incumbent on us to mitigate the risks and to help leadership understand how the organization is using AI, what processes and controls are in place, and how we are working with the development teams to ensure that they are implemented according to company rules.

STRATEGIES FOR SECURING AGENTIC AI

Security is everyone's job, and we will need to take a multilayered approach to achieve the desired results. There will be activities that are manual tasks, based on a consistent checklist and other activities that we can automate. The goal will be to establish a balance. Especially knowing that we will need some manual activities to dive deeper into findings from the automated tasks. The balance will shift over time as more activities can be automated.

The multilayered approach needs to include the traditional cybersecurity measures that have been implemented today as well as additional measures and protections for AI. This will include extending policy-based access controls, logging, monitoring, real-time alerts, and detection mechanisms for suspicious or malicious activities compared to a baseline and overall safety measures across the board. All of this needs to be carefully managed to address the challenges and risks.

By adopting these additional measures now, before solutions deploy to production, you will be ready to harness the power of agentic AI, enhance your security posture, and be able to protect against evolving threats.

The following six approaches should be added to your existing measures:

First, start with a gradual implementation. Identify the governance, security controls, and requirements and put those into place in down-level environments. Work with the development teams to understand where those measures may miss the mark and need to be augmented. Depending on your environment, if the development team is already moving forward, you may have to slipstream

measures into place and evolve the implementation over time.

Second, identify where existing and traditional cybersecurity measures need to be modified. There will be specialized protections for AI and more so for agentic AI.

Third, put in place, and require end-to-end monitoring. Ensure that teams are:

- Logging and monitoring communication (inputs and outputs) from the LLM as well as all communication to and from each agent.
- Include things such as a correlation ID, to be able to track a process through the complete life cycle instance. Each implementation will also have custom output which needs to be understood by the security teams.
- Identifying prompt injections, data leakage or unexpected behaviors.
- Restricting, or placing strict validations on, user-provided prompts or user-provided text in prompts.

As issues arise, you will need to be able to review decisions made by the agents and the associated workflow and review input and output to ensure the system is transparent. You will then have the information required if, and when, an audit needs to take place.

Fourth, look at LLM-specific threats. Put in place procedures to isolate agents from critical systems, limit the access the agent has to resources and evaluate and validate prompts before they get submitted to the LLM. Use tools such as OpenAI's Moderation API to evaluate prompts and responses to ensure content filtering.

In addition, as you review prompts and responses, use either the Evals Framework from OpenAI or the PromptFlow SDK from Microsoft. These tools provide the ability to evaluate the response with a set of ideal answers to compare against the response. You can either compare against specific benchmarks you create or when evaluating open-ended questions, you can have the model grade itself and provide statistics.

Also, review the top 10 list of threats published by OWASP. This will help you identify threats other than those we have covered here, and help you plan for mitigating those threats.

Fifth, incorporate automation frameworks that help testers, red team test groups, and security teams to proactively uncover risks. You should be performing a red team test and testing for responsible AI simultaneously. You will need to have a group dedicated to red-team activities and use a tool such as PyRIT. This tool helps to proactively uncover risks.

Red teaming an agentic AI system is different from traditional systems.

Agentic AI and traditional AI systems are nondeterministic, and scripts will need to be run multiple times. Each time the scripts are run the output will differ. You need to take this variability into account as you test each scenario. You also have to keep in mind that due to the agentic workflow logic, the LLM itself, the variability in prompts, and the agent behavior will result in more variability. You will also find that executing the same task against the same scenario will respond differently, and you will need to run more tests and test scenarios to cover any potential blind spots. Have your development teams create a map of all rules and flow possibilities through the process.

As with any tool, you won't be able to, and shouldn't always, automate everything. Use a tool such as PyRIT along with manual testing. Manual testing will allow testers to test specific trouble areas as well as perform deeper dives into any areas the automation testing uncovered.

Make sure you are also providing monitoring and logging of your automation tests. This will help test the process of tracing issues and the team as it dives in deeper with manual tests. Test the process of using the logged data to ensure transparency and auditability at this stage, instead of when an issue presents itself in production.

Sixth, work with other cybersecurity experts to compare measures and practices. Continue to build out your governance framework and always add and refine your procedures.

THE FUTURE OF AGENTIC AI: PROMISING AND FULL OF POSSIBILITIES

The wide range of benefits, capabilities, and efficiencies that can be offered to the business make this the perfect time to explore this technology.

However, the associated risks and security threats cannot be ignored. We must make sure that we are broadening the corporate culture so that security is everyone's responsibility. It is incumbent on teams to log all interactions, monitor the system, and ensure that there are human controls in place. Tools must be incorporated into the end-to-end processes, to proactively find issues before they erode user and business confidence. Transparency, human oversight, and AI safety must always be top of mind.

Security teams need to outline controls and governance, security measures, and rules. Development teams need to educate themselves, not only on these rules and requirements but also on the risks they will encounter and the mitigations they need to put in place. ■