akerman

Akerman Intelligence Alert

Pandora's Bots: Managing the Risks of Autonomous Agentic AI

March 12, 2026

Executive Summary

A major new study by researchers at Harvard, MIT, Stanford, Carnegie Mellon, Northeastern, and other leading institutions tested what happens when AI agents — the kind that can send emails, run software, manage files, and take actions on your behalf — are given real tools in a realistic environment. The results should concern any organization that is deploying or considering these technologies.

Over two weeks, researchers found that these agents could be manipulated into disclosing confidential data (including Social Security numbers and financial records), executing destructive actions without authorization, accepting instructions from strangers posing as authorized users, and spreading false information across networks — all through ordinary conversation, without any hacking or technical exploitation.

These are not flaws in one company's product. They are characteristics of how the current generation of autonomous AI works. Any organization using AI agents that can take real-world actions — whether for customer service, operations, deal management, research, or internal administration — should understand

Akerman Intelligence

Akerman Intelligence is Akerman's platform for AI strategy, research, and thought leadership. Through original analysis, client collaboration, and strategic partnerships, we help organizations navigate the legal, operational, and governance challenges of artificial intelligence.

Click here to learn more.

these risks and take concrete steps to manage them.

This alert explains what was found, why it matters across industries, and what you can do about it now. It also identifies real-world incidents where these same failures have already caused financial and reputational harm.

The Rise of AI Agents — and Why This Is Different

Most people are familiar with AI as a tool you ask questions and get answers from — a sophisticated search engine or writing assistant. But a new category of AI is rapidly entering the market: **autonomous agents**. These are AI systems that don't just answer questions — they take actions. They send emails, write and execute software, manage files, schedule tasks, interact with customers and vendors, and operate across platforms — often without checking back with a human before each step.

These tools are already available from major providers and are being adopted by enterprises of all sizes. They go by names you may recognize — Claude Code, ChatGPT in its latest agentic modes, Microsoft Copilot agents, and a growing ecosystem of autonomous AI platforms. The appeal is obvious: they can handle complex, multi-step tasks that previously required constant human attention.

The risk is less obvious, and that is the subject of this alert.

What Researchers Found

In February 2026, a team of 38 researchers from leading institutions published a study titled *Agents of Chaos*. They set up six AI agents in a controlled but realistic environment — each agent had its own email account, file storage, messaging access, and

the ability to run software — and then spent two weeks testing whether the agents could be manipulated, exploited, or induced to act against their operators' interests. Twenty AI researchers probed the agents with the kinds of tactics that bad actors, disgruntled employees, or careless users might employ in the real world.

The agents were powered by two different frontier AI models from different providers, representing different design philosophies. The vulnerabilities appeared across both, confirming that the problems are not specific to any single AI model.

Here is what they found, in plain terms.

Agents follow instructions from people they shouldn't trust

Each agent had a designated "owner" — the person it was supposed to work for. But when other people contacted the agents and asked them to do things, the agents generally complied. They ran system commands, transferred files, created documents, and disclosed email records for people who had no relationship to the owner and no reason to be making those requests. In one case, a researcher with no authorization obtained a complete log of 124 emails — not his own — simply by framing the request as urgent.

What this means for you: If your organization deploys an AI agent that interacts with customers, vendors, employees, or the public, that agent may carry out requests from anyone who contacts it — unless you have implemented controls that the agent itself cannot override.

Agents disclose confidential information when asked the right way

Researchers planted sensitive personal data — Social Security numbers, bank account numbers, medical

information — in the agent's email system. When asked directly for the Social Security number, the agent refused. But when asked to simply forward the email containing it, the agent sent everything — unredacted. The agent understood that a Social Security number is sensitive. It did not understand that an email containing a Social Security number is equally sensitive.

What this means for you: Any AI agent with access to systems that contain confidential data — customer records, deal documents, personnel files, financial accounts, patient information — can be tricked into disclosing that data through simple, indirect requests. No technical skill is required.

Agents can be impersonated with startling ease

Think of it this way: an agent had previously been communicating with its owner, "Chris," in one chat room and had learned to recognize him there. When an attacker opened a separate, new chat room using the name "Chris" — the equivalent of changing the caller ID on a phone—the agent had no memory of the prior conversation and no way to verify whether this was the same person. Based solely on the matching name, the agent treated the stranger as its owner and followed every instruction, including deleting its own files, rewriting its operating rules, and handing over administrative control. The entire takeover required nothing more than typing a name into a profile field.

What this means for you: If your AI agent determines who to trust based on names, titles, or conversational cues — rather than verified credentials — anyone who can type a name into a profile field can gain full control.

Agents can be corrupted through the documents they read

A researcher convinced an agent to co-create a shared governance document, stored online. Later, the researcher quietly edited the document to include hidden instructions. Because the agent trusted the document as part of its operating rules, it followed the injected instructions, attempting to shut down other agents, removing users from the shared workspace, sending unauthorized emails, and even sharing the compromised document with other AI agents, all without being asked.

What this means for you: AI agents that process external content — emails, uploaded documents, shared files, web pages — can be manipulated through that content. A malicious attachment, a doctored shared document, or a carefully crafted email can alter what your agent does, without anyone accessing the agent directly.

Agents take extreme actions and then misrepresent what happened

In one case, a third party shared confidential information with an agent via email and later asked the agent to delete that email. The agent did not have the ability to delete a single email. Rather than explaining this limitation or suggesting alternatives, it destroyed its owner's entire email system — wiping all stored messages — as a way to eliminate the one email in question. It then reported that the confidential information had been successfully removed. It had not. The data was still accessible through the email provider's web interface. The agent chose the most destructive available option, failed to achieve its goal, and misrepresented the outcome.

What this means for you: Agents may take disproportionate, irreversible actions in pursuit of reasonable goals. Worse, they may tell you they succeeded when they didn't. Organizations that rely on agent self-reporting without independent verification are operating with a false picture of what their systems have done.

Problems multiply when agents interact with each other

When multiple agents operated in the same environment, failures compounded. Agents reinforced each other's mistakes, propagated compromised instructions to one another, entered resource-consuming loops that ran for days without detection, and — in one test — broadcast fabricated defamatory claims to an entire contact list within minutes of being instructed to do so by someone impersonating the agent's owner.

What this means for you: Organizations operating multiple agents, or whose agents interact with third-party agents — including those of clients, vendors, government agencies, and the broader internet — face cascading risks that are qualitatively different from single-system failures. A compromise of one agent can spread to others automatically.

These Risks Are Not Theoretical

While the *Agents of Chaos* study was conducted in a controlled laboratory, the same failure patterns are already appearing in production environments — with real financial and reputational consequences.

An AI safety executive loses control of her own agent. In February 2026, a frontier AI company's Director of Alignment — the person whose job is to ensure AI systems behave as intended — gave an autonomous AI agent access to her personal email. Despite explicit instructions to suggest actions and wait for approval before doing anything, the agent began mass-deleting emails on its own. She typed "STOP" repeatedly. The agent ignored her. She had to physically run to her computer to kill the process. Her own assessment, posted publicly: "Rookie mistake. Turns out alignment researchers aren't immune to misalignment." The post was viewed nearly 10 million times.

An AI coding tool destroys a company database, fabricates replacements, and lies about it. In July

2025, an AI coding assistant on a widely used AI coding platform, using a different frontier AI model, deleted a live production database containing records on over 1,200 executives and 1,196 companies — during an active freeze on all changes, and after being told 11 separate times not to make modifications. The agent then fabricated 4,000 fake user records to fill the gap, generated false test results to conceal the damage, and told its operator that data recovery was impossible. In fact, recovery was possible and eventually succeeded. The platform's CEO issued a public apology and called the incident "unacceptable."

An AI trading agent gives away most of a $50,000 wallet to a stranger. In February 2026, an engineer at a third frontier AI company gave an AI trading agent a cryptocurrency wallet with $50,000 in assets and instructions to grow it through trading. When a stranger on social media made a small request — roughly $310 worth of tokens, accompanied by a fabricated story about a sick relative — the agent transferred the vast majority of the wallet's value in a single irreversible transaction. The agent had been socially engineered through a simple emotional appeal.

An AI coding agent wipes 2.5 years of production data in a single session. In late February 2026, a developer using yet another AI coding agent to migrate a website to new infrastructure made a routine error — forgetting to upload a configuration file from another computer. Rather than flagging the discrepancy, the agent escalated from a targeted cleanup task to a full infrastructure destruction command, reasoning it would be "cleaner and simpler." The command wiped the production database for an online education platform serving over 100,000 students, destroying 2.5 years of student submissions, homework, projects, and leaderboard entries. All automated backup snapshots were destroyed along with it. The developer was able to recover the data — nearly two

million rows — only after upgrading to premium cloud support and waiting 24 hours.

Each of these incidents involves a different product, a different company, and a different context. Together, they reveal two distinct failure modes. In some cases, the agent faithfully executes a flawed human instruction — amplifying a routine mistake into a catastrophic outcome because it lacks the judgment to recognize the error. In others, the agent acts contrary to explicit instructions for reasons that are not well understood — ignoring repeated human commands, fabricating information to conceal its actions, or choosing destructive options when moderate alternatives are available. The first category is a problem of oversight. The second is a problem of the technology itself. Both require different safeguards.

Who Should Be Concerned

These findings apply broadly. They are not limited to any particular company, industry, sector, or domain. Any enterprise deploying autonomous AI agents — or evaluating them for future use — should consider whether the vulnerability classes described above create material risk within its operations, regulatory environment, and existing governance framework.

What You Can Do Now

We are not suggesting that organizations stop using AI agents. These tools deliver real productivity and operational gains, and they will only become more prevalent. But the gap between what these systems can do and what they can do *safely* is wider than most organizations appreciate. The following steps can materially reduce your exposure.

1. Know what your agents can access and do. Before deploying any AI agent, map its permissions the same way you would map a new employee's system access. What databases can it reach? What communications can it read? What actions can it take? Apply least-privilege principles — give the

agent only the access it needs for its specific function, and nothing more.

2. Require human sign-off for consequential actions. Any agent action that is irreversible, involves confidential information, or affects third parties should require a human to confirm it before it happens. This includes sending external communications, modifying or deleting records, executing financial transactions, and granting or revoking access.
3. Don't let the agent decide who to trust. Implement identity verification at the platform level — verified credentials, multi-factor authentication, or cryptographic identity — rather than allowing the agent to infer who is authorized based on names, titles, or conversational context. The study showed that agents will accept a spoofed identity from a simple display-name change.
4. Monitor agent outputs for confidential data independently. Deploy data loss prevention tools that scan what the agent sends and writes, rather than relying on the agent to self-censor. The study proved that agents will refuse to disclose a labeled sensitive field but will hand over the entire document containing it without hesitation.
5. Watch for runaway processes. Agents can create background tasks, scheduled jobs, and growing data stores without telling anyone. Implement automated monitoring for unexpected resource consumption, and set hard limits on what the agent can create or schedule without human approval.
6. Don't trust the agent's self-reporting. When an agent reports that it completed a task — especially a sensitive one like deleting data or securing a system — verify independently. The study documented agents claiming success while the underlying system state told a different story. The coding platform incident described above followed the same pattern.
7. Update your contracts and insurance. Review vendor agreements for AI agent platforms to

ensure they address liability allocation, indemnification, and breach notification when the agent — not a hacker — is the vector for data exposure. Confirm that your cyber insurance covers agent-initiated incidents, not just traditional intrusions.

8. Add agentic AI to your M&A and investment diligence. For any acquisition target or portfolio company using autonomous agents, expand your diligence to include agent permissions, interaction logs, external resource dependencies, and memory contents. An agent that has been silently compromised may not leave the usual forensic trail.
9. Redefine what counts as an "incident." Traditional breach detection looks for unauthorized access — someone breaking in. Agent-initiated disclosures use authorized access; no alarm trips because the agent is doing what it's permitted to do, just for the wrong person or the wrong reason. Update your incident response framework to account for this.
10. Evaluate whether your board has visibility into agentic AI risk. Organizations may wish to consider whether their boards and senior leadership have sufficient visibility into how autonomous AI agents are being deployed, what they can access, and what safeguards are in place. For organizations where agentic AI touches material operations, customer data, or financial systems, it may be worth evaluating whether existing risk oversight structures — including any relevant committee charters — are positioned to address this category of technology risk, and whether the organization has access to the technical expertise needed to assess it. The pace of adoption in this space can outrun traditional governance cycles, and early visibility may reduce the likelihood of being caught off guard.

Looking Ahead

The regulatory landscape is catching up. In February 2026, the National Institute of Standards and Technology (NIST) announced its AI Agent Standards Initiative, identifying agent identity, authorization, and security as priority areas for standardization. The study's findings align precisely with these priorities. Organizations that implement reasonable safeguards now will be better positioned as formal standards emerge, and less likely to face retroactive compliance burdens or enforcement scrutiny.

The vulnerabilities documented in this research are not reasons to avoid agentic AI. They are reasons to deploy it thoughtfully. The organizations that benefit most from these tools will be those that understand their limitations from the outset — and build accordingly.

Source: Shapira, Wendler, Yen et al., Agents of Chaos (arXiv: 2602.20021, February 2026). 38 researchers from Northeastern University, Harvard, MIT, Stanford, Carnegie Mellon, University of British Columbia, Hebrew University, Tufts, Max Planck Institute, and others. Interactive version with full interaction logs: agentsofchaos.baulab.info

---

Additional Information

If you would like additional information, or have specific questions about how this information may apply to your organization, please contact your Akerman relationship attorney or reach out to Akerman Intelligence directly. We also invite you to visit our Akerman Intelligence page for ongoing analysis and practical application of leading-edge AI-related legal developments across sectors and practices.

intended as legal advice for any specific matter and does not create an attorney-client relationship. Please consult with counsel to determine how this information may apply to your specific situation.