CSO

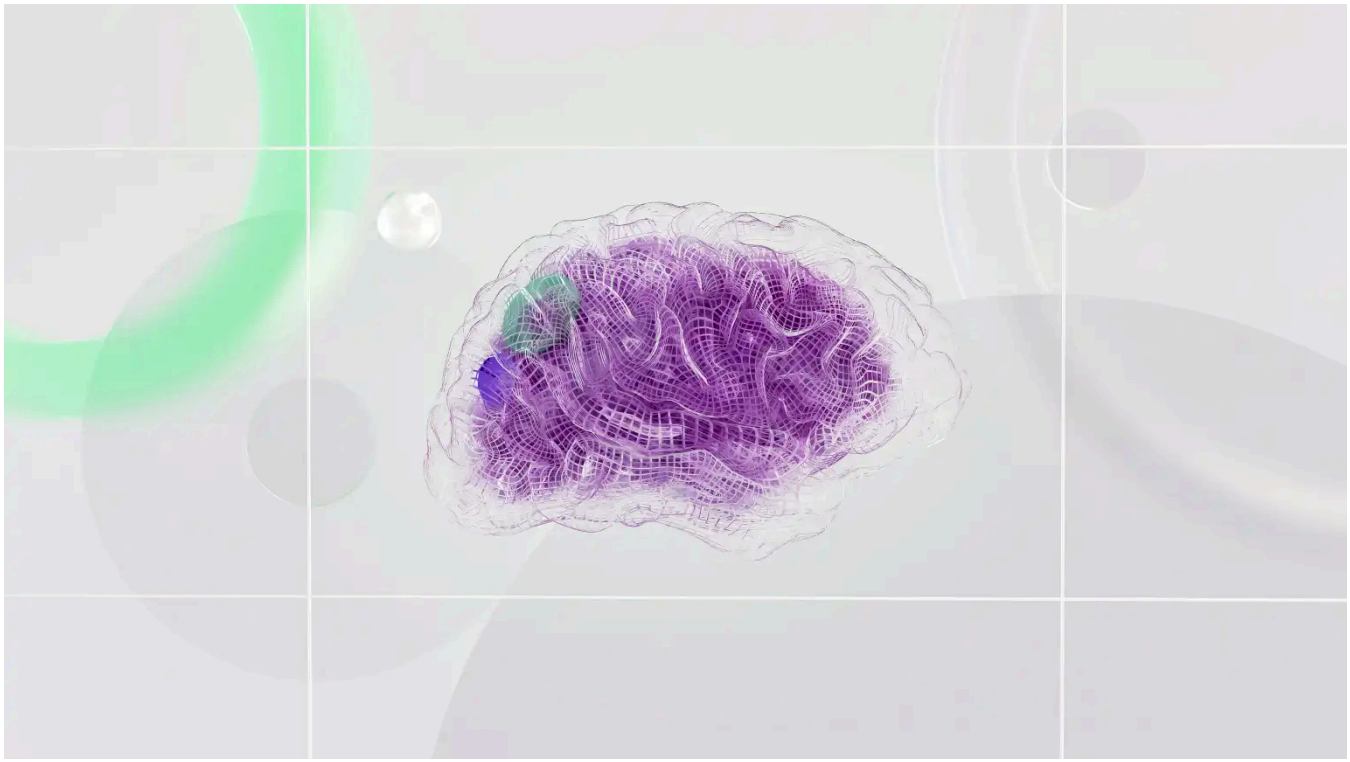
by **Linda Rosencrance**
Contributing Writer

9 ways CISOs can combat AI hallucinations

Feature

Apr 1, 2026 • 9 mins

AI-based compliance assessment tools might not be ready for fully independent assessments, if CISOs are using these tools we share some best practices to ensure accuracy and avoid risks or fines.



Credit: Google Deepmind

AI hallucinations are a well-known problem and, when it comes to compliance assessments, these convincing but inaccurate assessments can cause real damage with

poor risk assessments, incorrect policy guidance, or even inaccurate incident reports.

Cybersecurity leaders say the real trouble starts when AI moves past writing summaries and begins making judgment calls. That's when it's asked to decide things such as whether security controls are doing their job, if a company is meeting compliance standards, or if an incident was handled the right way.

Here are nine ways CISOs can tackle the problem of AI hallucinations.

Keep humans in the loop for high-stakes decisions

Fred Kwong, vice president and CISO at DeVry University, says his team is carefully testing AI in governance, risk, and compliance work, especially in third-party risk assessments. He notes that while AI helps review vendor questionnaires and supporting evidence that assess the security posture of those vendors, it doesn't replace people.

"What we're seeing is the interpretation is not as good as I would want it to be, or it's different than how we're interpreting it as humans," Kwong says.

He explains that AI often reads control requirements differently than experienced security professionals do. Because of that, his team still reviews the results manually. For now, AI is not saving much time because the trust in the technology just is not there yet, he says.

Mignona Coté, senior vice president and CISO at Infor, agrees that human oversight is critical, especially in risk scoring, control assessments, and incident triage. "Keep the human in the loop, full stop," says Coté, who sees AI as a productivity tool, not something that should make final decisions on its own.

Treat AI outputs as drafts, not finished products

One of the biggest risks is over-trusting AI, according to security experts. Coté says her organization changed its policy so AI-generated content cannot go straight into compliance documentation without a human review.

"The moment your team starts treating an AI-generated answer as a finished work product, you have a problem," she says. "Treat every output as a first draft as opposed to a final one. There will come a point where repetitive questions will have repetitive answers. By labeling those answers and time stamping them at origination time, they can be addressed at scale."

Srikumar Ramanathan, chief solutions officer at Mphasis, says this over-trust often comes from what he calls "automation bias." People naturally assume that something written clearly and confidently must be correct.

To counter that, he says companies need to build an “active skepticism” culture. “[That means] looking upon AI outputs as unverified drafts that require a signature of human accountability before they are actionable,” he explains.

Demand proof, not polished prose, from vendors

When vendors say their AI can “assess compliance” or “validate controls,” security leaders say buyers need to ask the tough questions.

Kwong says he pushes vendors to provide traceability of the answers that the AI gives so his team can see how the AI reached its conclusions. “Without that traceability, it makes it even that much harder for us to identify,” he says.

Ramanathan says buyers should ask whether the system can point to the exact evidence behind its answer, such as a time-stamped log entry or a specific configuration file. If it can’t, the tool may just be generating text that sounds right.

Puneet Bhatnagar, a cybersecurity and identity leader, says the key question is whether the AI is actually analyzing live operational data or just summarizing documents. “If a vendor cannot show a deterministic evidence path behind its conclusion, it’s likely generating narrative – not performing an assessment,” says Bhatnagar who most recently

served as SVP and head of identity management at Blackstone. “Compliance isn’t about language. It’s about proof.”

Stress-test models before extending trust

Kwong recommends testing AI tools to see how consistent they are. For example, send the same data through twice and compare the results.

“If you send the same data again, is it spitting back the same result?” he asks.

If answers change significantly, that’s a red flag. He also suggests removing important evidence to see how the model reacts. If it confidently gives an answer anyway, that could signal a hallucination.

Coté says her team checks AI outputs against other tools, including scanning systems and external penetration testing results. “And we don’t extend trust to any AI tool until it has proven itself against known outcomes repeatedly,” she says.

Measure hallucination rates and monitor drift

Security leaders say organizations need to track how accurate AI is over time. Kwong says teams should regularly compare AI-generated assessments with human reviews and study the differences. That process should happen at least quarterly.

Ramanathan suggests tracking metrics such as “drift rate,” which measures how often AI conclusions differ from human reviews. “A model that was 92% accurate six months ago and is 85% accurate today is more dangerous than one that’s been consistently at 80% because your team’s trust was calibrated to the higher number,” he notes.

He also recommends measuring how often cited evidence truly supports the AI’s claims. If hallucination rates climb too high, organizations should reduce how much authority the AI has, for example, downgrading it to a less autonomous role in their governance models.

Watch for contextual blind spots in compliance mapping

Bhatnagar says the most dangerous hallucinations happen when AI is asked to make judgment calls about control effectiveness, regulatory gaps, or incident impact.

AI can produce what he calls “plausible compliance”, or answers that sound convincing but are wrong because they lack real-world context. Compliance often depends on technical details, compensating controls, and operational realities that documentation alone doesn’t show.

Ramanathan adds that AI often struggles with the nuance of permissive language, (“may,” “can”) versus restrictive language (“must,” “is required to”).

“For example, AI often misinterprets permissive language like ’employees may access the system after completing training’ as a strict, enforceable rule, treating optional permissions as mandatory controls,” Ramanathan explains. “This causes AI to overestimate the authority of permissive or vague language, resulting in incorrect assumptions about whether policies are properly enforced or security measures are effective.”

Push back on generic or identical assessments

Some vendors overstate what their AI tools actually do. Bhatnagar says many tools summarize documents or generate gap reports but vendors market those features as if they’re doing full, automated compliance checks.
The risk increases when multiple customers receive nearly identical assessments. Organizations may believe their controls were thoroughly evaluated when the AI only performed a surface-level document review.

Ramanathan says this creates false confidence and broader industry risk. If one popular model has a flaw, that blind spot can spread widely.

Bhatnagar adds that he has seen vendors market AI tools as assessing whether organizations are compliant, even when multiple customers receive structurally similar or nearly identical assessments.
In those situations, the tool may not actually be analyzing company-specific policies or evidence but instead generating text that appears customized without being grounded in reality, he says. “We are still in the early stages of separating AI narrative generation from AI-based verification,” he says. “That distinction will define the next phase of governance tooling.”

Reinforce accountability in audits and legal reviews

From a regulatory standpoint, AI does not remove responsibility, according to experts. Ramanathan says regulators are clear that duty of care stays with corporate officers.

“If an AI-generated assessment misses a material weakness, the organization is liable for ‘failure to supervise,'” he says. “We are already in an era wherein relying on unverified AI outputs could be seen as gross negligence. If your audit findings are wrong because of an AI error, you haven’t just failed an audit, you are held responsible for filing a misleading regulatory statement. ‘AI told me so’ is not a defense.”
Coté says being able to show that a human reviewed and approved each consequential decision is critical during audits. “The key is proving a human was at every consequential

decision point, with a timestamp and an audit trail to back it up," she notes.

Be cautious with automated regulatory mapping

Ramanathan says that one of the biggest compliance risks appears when companies rely on AI to automatically map internal controls to regulatory frameworks, such as GDPR or SOC 2.

"The greatest compliance risk by far is in automated regulatory mapping," he notes. "The AI might confidently claim a control exists or satisfies a requirement based on a linguistic pattern rather than a functional or operational reality."
For example, an AI tool might see an encryption setting listed in a database configuration and assume encryption is active, even if that feature is turned off in the system.

Ramanathan says this can create "a massive security gap where a company believes they are audit-ready, only to discover during a breach that their AI-verified defenses were nonexistent or misconfigured."

To reduce that risk, he says organizations need to structure their policies and regulations more clearly and connect them to enforceable technical rules rather than relying only on AI to interpret documents.

Artificial Intelligence[https://www.csoonline.com/artificial-intelligence/] Compliance[https://www.csoonline.com/compliance/] Risk Management[https://www.csoonline.com/risk-management/] Security[https://www.csoonline.com/security/]

SUBSCRIBE