IMPACT FUND
Strategic Litigation for Social Justice

Protocol for the Use of Generative Artificial Intelligence in Operations and Program Delivery

Effective Date: August 1, 2025

1. Statement of Purpose

This Protocol is intended to help Impact Fund employees use approved generative artificial intelligence ("AI") tools with thoughtfulness and care to safely increase operational and program effectiveness.

Generative AI tools allow users to enter prompts to receive human-like text, images, sound, or videos created by artificial intelligence. They can also summarize content, categorize information, and perform numerous other tasks. Examples include ChatGPT, Google's Bard, and Microsoft Bing.

2. Factual Background

The field of generative AI is changing rapidly. The following facts are fundamental to this Protocol:

- Generative AI tools are prone to "hallucinations," the creation of false answers or information, and the use of stale information.

- Generative AI tools rely on patterns of information and can replicate preexisting bias, including implicit and explicit biases that further systems of oppression that the Impact Fund is affirmatively working to eliminate.

- Information entered into publicly available generative AI tools is shared publicly, which may violate confidentiality, attorney-client privilege, and work product protections, as set forth in the Impact Fund's Information Security Protocol and Employee Policy Handbook.

- Early studies indicate that overreliance on generative AI tools impedes learning, critical thinking, creativity, analysis, information retention, skill development, and independent problem solving. Routine use decreases self-confidence and increases reliance on generative AI tools.

3. Guiding Principles

The Impact Fund adopts the following foundational principles to guide its use of generative AI tools for organizational purposes:

a. Use of generative AI tools at the Impact Fund will be human centered. Our work is rooted in authentic human ingenuity and connection, therefore our colleagues and constituents must understand that all Impact Fund communications and work product come from a person. Misuse of generative AI tools can damage our internal trust in each other and external trust in our work and professional reputation.

b. Generative AI tools are not a substitute for human judgment and creativity.

c. Employees hold ultimate responsibility for producing reliable work product via replicable means, even when using generative AI tools.

4. Acceptable Use of Generative AI Tools

DO:

a. Use only generative AI tools that have been approved by the Impact Fund for use.

b. Be familiar with the generative AI tool's policies on data retention, data sharing, and self-learning.

c. Undertake training provided by the Impact Fund on proper use of generative AI tools.

d. Keep your team lead informed on the ways you use generative AI tools in your day-to-day work and obtain advance approval before using generative AI tools for specific projects. Team leads may also specify tasks or projects that are inappropriate for the use of generative AI tools.

e. Review and verify all generative AI outputs before use, including by identifying and reviewing primary sources and by fact-checking AI outputs against independent research.

f. Review generative AI outputs for algorithmic (inherited) discrimination, including the perpetuation or recreation of bias.

g. Maintain the privacy of the Impact Fund's intellectual property and its Confidential, Restricted, or Privileged/Protected Information (as defined by our Information Security Protocol) at all times.

h. Comply with all Impact Fund policies, applicable laws, and vendor terms of use.

DO NOT:

a. Use any generative AI tool for any purpose that could (1) negatively impact the trust that our clients, grantees, and constituents have in us, or (2) endanger the authenticity or reliability of our communications or work product.

b. Represent work generated by a generative AI tool as your own original work.

c. Place the Impact Fund's intellectual property or its Confidential, Restricted, or Privileged/Protected Information (as defined by our Information Security Protocol) into any generative AI tool.

d. Use generative AI tools for legal or other substantive research without reviewing all primary sources and independently confirming the accuracy of outputs.

e. Use "self-learning" generative AI tools, unless they have been modified to not store or repurpose user-provided data.

f. Use generative AI tools to attempt to re-identify anonymized data.

g. Use generative AI tools to create visual or written content that replaces the ordinary hire and compensation of artists.

h. Violate the intellectual property rights of authors or visual artists by requesting outputs in their specific style.

i. Use generative AI tools provided by the Impact Fund for personal gain, unlawful activities, or any purpose that could damage the Impact Fund or our community.

j. Integrate any generative AI tool with internal software without specific written permission from your team lead and the Deputy Director.

5. Generative AI Tools Approved for Use

Employees may only use generative AI tools previously approved by the organization. To be approved, technologies must protect data from public disclosure and unauthorized access, including through anonymization, encryption, and secure storage practices.

A list of approved technologies is available upon request. Employees are invited to recommend new tools to the Deputy Director for approval.

6. Violations

Maintaining this Protocol is considered part of each employee's job duties. Violating this Protocol may result in disciplinary action as described in Section 2.15 ("Performance Improvement & Development") of the Employee Policy Handbook. Disciplinary action may include suspension, access restrictions, work assignment limitations, or more severe penalties up to and including immediate termination, in accordance with applicable law. Suspected violations of this Protocol should be reported to the HR Director or the Executive Director.

7. Disclosure

The Impact Fund is transparent in its use of generative AI tools internally and externally to encourage trust in our work. A copy of this Protocol is publicly available on our website, www.impactfund.org.

# # #

Last updated on January 15, 2026

Vigilante Lawyers Expose the Rising Tide of A.I. Slop in Court Filings

More lawyers are using artificial intelligence to write legal briefs. Some colleagues are publicizing the A.I.-generated errors.

By Evan Gorelick

Nov. 7, 2025

Earlier this year, a lawyer filed a motion in a Texas bankruptcy court that cited a 1985 case called Brasher v. Stewart.

Only the case doesn't exist. Artificial intelligence had concocted that citation, along with 31 others. A judge blasted the lawyer in an opinion, referring him to the state bar's disciplinary committee and mandating six hours of A.I. training.

That filing was spotted by Robert Freund, a Los Angeles-based lawyer, who fed it to an online database that tracks legal A.I. misuse globally.

Mr. Freund is part of a growing network of lawyers who track down A.I. abuses committed by their peers, collecting the most egregious examples and posting them online. The group hopes that by tracking down the A.I. slop, it can help draw attention to the problem and put an end to it.

While judges and bar associations generally agree that it's fine for lawyers to use chatbots for research, they must still ensure their filings are accurate.

But as the technology has taken off, so has misuse. Chatbots frequently make things up, and judges are finding more and more fake case law citations, which are then rounded up by the legal vigilantes.

“These cases are damaging the reputation of the bar,” said Stephen Gillers, an ethics professor at New York University School of Law. “Lawyers everywhere should be ashamed of what members of their profession are doing.”

Since the introduction of ChatGPT in 2022, professionals in fields from medicine to engineering to marketing have wrestled with how and when to use chatbots. Many companies are experimenting with the technology, which can come tailored for workplace use.

For lawyers, a federal judge in New York helped set the standard when he wrote in 2023 that “there is nothing inherently improper” about using A.I., although they must check its work. The American Bar Association agreed, adding that lawyers “have a duty of competence.”

Still, according to court filings and interviews with lawyers and scholars, the legal profession in recent months has increasingly become a hotbed for A.I. blunders. Some of those stem from people’s use of chatbots in lieu of hiring a lawyer. Chatbots, for all their pitfalls, can help those representing themselves “speak in a language that judges will understand,” said Jesse Schaefer, a North Carolina-based lawyer who contributes cases to the same database as Mr. Freund.

But an increasing number of cases originate among legal professionals, and courts are starting to map out punishments of small fines and other discipline.

The problem, though, keeps getting worse.

That’s why Damien Charlotin, a lawyer and researcher in France, started an online database in April to track it.

Initially he found three or four examples a month. Now he often receives that many in a day.

Many lawyers, including Mr. Freund and Mr. Schaefer, have helped him document 509 cases so far. They use legal tools like LexisNexis for notifications on keywords like “artificial intelligence,” “fabricated cases” and “nonexistent cases.”

Some of the filings include fake quotes from real cases, or cite real cases that are irrelevant to their arguments. The legal vigilantes uncover them by finding judges' opinions scolding lawyers.

Peter Henderson, a Princeton computer science professor who started his own A.I. legal misuse database, said his lab was working on ways to find fake citations directly rather than relying on hit-or-miss keyword searches.

The lawyers say they don't intend to shame or harass their peers. Mr. Charlotin said he avoided prominently displaying the offenders' names for that reason.

But Mr. Freund said a benefit of a public catalog was that anyone could see whom they "might want to avoid."

And in most cases, Mr. Charlotin added, "the attorneys are not very good."

Eugene Volokh, a law professor at the University of California, Los Angeles, blogs about A.I. misuse on The Volokh Conspiracy. He has written about the issue more than 70 times, and contributes to Mr. Charlotin's database.

"I like sharing with my readers little stories like this," Mr. Volokh said, "stories of human folly."

One involved Tyrone Blackburn, a New York lawyer focusing on employment and discrimination, who used A.I. to write legal briefs that contained numerous hallucinations.

At first he thought the defense's allegations were bogus, Mr. Blackburn said in an interview. "It was an oversight on my part."

He eventually admitted to the errors and was fined $5,000 by the judge.

Mr. Blackburn said he had been using a new legal A.I. tool and hadn't realized it could fabricate cases. His client, who he was representing for free, fired him and filed a complaint with the bar, Mr. Blackburn added.

(In an unrelated matter, a New York grand jury indicted Mr. Blackburn last month on allegations he rammed his car into a man trying to serve him legal documents. Attempts to reach Mr. Blackburn for additional comment failed.)

Court-ordered penalties “are not having a deterrent effect,” said Mr. Freund, who has publicly flagged more than four dozen examples this year. “The proof is that it continues to happen.”

Evan Gorelick is a New York-based writer for The Morning, the flagship daily newsletter of The Times.

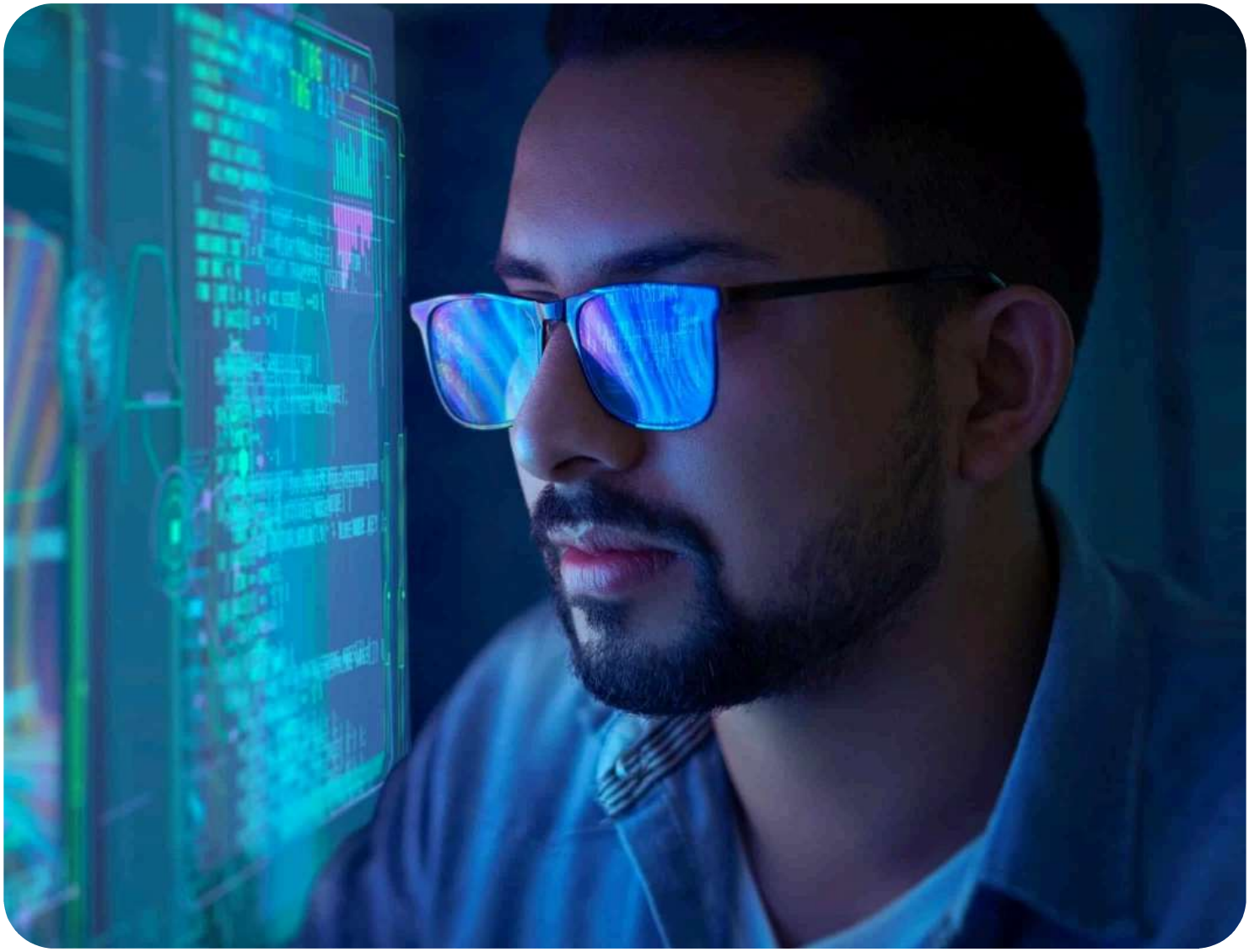
A version of this article appears in print on , Section B, Page 3 of the New York edition with the headline: Vigilante Lawyers Expose A.I. Slop in Court

RESPONSIBLE AI

You Won't Get GenAI Right If You Get Human Oversight Wrong

By Steven Mills, Noah Broestl, and Anne Kleppe

ARTICLE MARCH 27, 2025 12 MIN READ



soufflés in the oven and the weather on Mount Everest, generative AI (GenAI) calls for vigilance. Companies get that, but their precautions aren't always preventive. Many organizations assume that humans in the loop will catch any problems and, fail-safe in place, they'll deploy GenAI carefree. Yet while human oversight is crucial for mitigating the risks of GenAI, it's still only one part of a solution. And the typical approach—assigning people to review output—carries risks of its own. The problem is that organizations rely on human oversight without *designing* human oversight. They have good intent but lack a good model.

That model isn't elusive. But it has several components that must be designed alongside the GenAI system.

Human oversight works best—which means it actually works—when it is combined with system design elements and processes that make it easier for people to identify and escalate potential problems. It also needs to be paired with other key ingredients of GenAI vigilance, including testing and evaluation, clear articulation of use cases (to ensure that GenAI systems don't deviate from their intended use), and response planning. Getting this right means thinking about human oversight at the product conception and design stage, when organizations are building a business case for a GenAI solution. Tacking it on during implementation—or worse, just prior to deployment—is too late.

> "Human oversight works best—which means it actually works—when combined with system design elements and processes that make it easier for people to identify and escalate problems.

A Fail-Safe That Often Fails

One of the unique traits of GenAI is that it can err in the same way humans err: by creating offensive content, demonstrating bias, and exposing sensitive data, for instance. So having humans check the output would seem a logical countermeasure. But there are a number of reasons why simply putting a human in the loop isn't the fail-safe that organizations envision.

- Automation Bias. In effect, success breeds complacency. Humans will review initial outputs, see no errors, and quickly come to trust the system. Appraisals becomes cursory or even nonexistent. Consider something as commonplace as GPS navigation. A driver may have entered addresses hundreds, even thousands of times and in each case, the system successfully directed them to the destination. So when the system takes them on an unpaved path, without an expected landmark in sight, there's a natural inclination to trust that the technology "knows what it's doing" and keep taking those turns. We've all heard the stories: drivers navigating to the water's edge or past it; trips that go on and on. Often there's a simple explanation, such as an address that matches multiple locations (think 15 Main Street) or road work not yet reflected in the software. But the system's track record created a hyperconfidence in its capabilities. And vital interventions never happened.

- Missing Context. GenAI systems often produce output without any additional information, such as supporting evidence. This lack of context can make it hard for reviewers to determine whether the answer is accurate or appropriate. As a result, reviewers face two choices: conduct additional research, cancelling out any efficiency gains from the system, or, more likely, accept the output at face value if it seems generally correct. Evaluation based on vibes rather than facts isn't a viable risk mitigation approach.

- Lack of Counterevidence. Even when systems provide supporting evidence, that information justifies only why the output is correct. Few systems also present counterfactual evidence. So, while reviewers see the case in favor of the output, they don't see the case against the output. Consider a GenAI solution that reviews whether permit applications are complete. The system produces a "yes" response because it sees that an application and the two required supporting documents have been filed. What the system should also share is that one of the supporting documents may be incomplete.

- A Disincentive Structure. GenAI is often employed to drive efficiency. Systems like co-pilots, chatbots, and customer self-service are all about simplifying processes and boosting productivity. But thoroughly evaluating GenAI output takes time—in many cases more than the system's designers envisioned when they set efficiency targets. Managers are often held to these targets, creating pressure on teams, intended or not, to keep the efficiencies coming. Concerned about the negative repercussions of slowing things down, people are likely to perform only cursory reviews of system outputs.

- Escalation Roadblocks. Many GenAI systems lack mechanisms to flag responses that the user believes are incorrect, leading to uncertainty about what to do next. But an even bigger problem, perhaps, is the

general skepticism that often permeates the review process. There's an assumption that the system is right and anyone claiming otherwise must better have a rock-solid case. Reviewers often need to take tedious administrative steps to justify their belief that the response is wrong. And that's assuming an escalation process even exists. Some organizations may have a policy requiring that users accept the output.

- **Misunderstanding GenAI Capabilities**. Knowledge about how GenAI works—its capabilities and limits—can vary wildly, even within a single organization. Hype, combined with the evolving nature of GenAI, often skews perceptions. As a result, many users view it as an almost magical technology and will second-guess themselves before questioning the GenAI system when they see a result that doesn't seem quite right.
- **Focusing on Accuracy but Not Scope**. More than with other technologies, GenAI can be taken "off track" if designers focus on what the system should do but not also on what it should not do. Output might be technically correct yet still "bad" if the system deviates from its intended use. Reviewers, however, aren't always briefed on use case boundaries. Their focus is on accuracy. So out-of-scope responses often go unquestioned.

> Evaluation based on vibes rather than facts isn't a viable risk mitigation approach.

Human Oversight by Design

Together, these factors paint a pretty grim picture of human oversight. And that's the point: simply telling people to watch over the AI isn't a solution. Eventually, a problem will go unnoticed—and then attract all too much notice. When that happens, the "we had human oversight" defense isn't going to fly with shareholders, customers, or the news crews camped outside the office.

Don't count on it flying with regulators or the courts, either. In December 2023, the Court of Justice of the European Union issued an opinion in a case related to assessing the creditworthiness of individuals. The court found that decisions to approve or deny credit applications, ostensibly made by humans, were effectively automated, as the humans routinely relied solely on algorithm-generated scores. This was a textbook case of automation bias. Put simply: the oversight was meaningless; the score was all that mattered.

Meaningful oversight requires more than putting humans in the loop. Companies need to treat oversight as an integral part of GenAI, not an add-on. They need to integrate it into the system's design and surrounding business processes and develop the procedures and organizational culture that enable people to identify problems—and do something about them. This may seem an onerous task, but in our experience, some best practices can guide the way.

> Meaningful oversight requires more than putting humans in the loop. Companies need to treat oversight as an integral part of GenAI, not an add-on.

Define a process around human oversight. Guidelines are better than vibes. Without a structured rubric to evaluate system outputs, human reviewers are often left to rely on hunches and intuition. That may work well for detectives on British television, but it's less than ideal for GenAI oversight. What should humans be looking for in the output? What are the red flags to consider? The idea is to develop a cookbook for the person interacting with the system that shows them, step by step, how to thoughtfully evaluate results. Like a recipe, the process is well defined and can be performed in a consistent way from one person to the next.

It's also important to specify, clearly and precisely, reviewer qualifications. Reviewers should have experience that's relevant to the output. For example, for a system that simplifies insurance claims processing—which includes technical tasks like assessing repair and replacement estimates—a skilled claims adjuster should handle the review.

Human reviewers also need an effective way to escalate errors. Simplicity works best. Companies should design steps that not only enable reporting and response but spark and accelerate it. In our experience, organizations that successfully scale AI devote 70% of their effort to people and processes. For human oversight, this means identifying issues that may hinder error escalation, whether it's metrics, policies, incentives, or all of the above. And it means designing GenAI solutions in a human-centered way, engaging with users to create features that facilitate output review, such as a "report" button built into the user interface. Finally, the process should include a roadmap for how the organization will respond, in terms of escalation, remediation, and communication, when a reviewer flags a potential GenAI failure.

Design systems to give evidence for and against outputs. GenAI development teams are valuable partners in human oversight. By considering how to drive the review process as they make design decisions, developers can improve the accuracy and efficiency of evaluations. We've found that one of the most powerful enablers is context. To that end, GenAI systems should generate a simple summary of both "for" and "against" evidence, giving reviewers a clearer way to decide whether to accept or flag output. There's an added bonus with this approach: when users ask a GenAI system to make the case, pro and con, for its response, the quality of that response tends to improve.

Track response rejection rates. Human oversight shouldn't work in isolation but rather as part of a holistic risk mitigation solution. A key component of this integrated approach is a comprehensive test-and-evaluation (T&E) process, one that leverages the strengths of both humans and automation. Robust T&E gives product development teams a good view of a system's accuracy. Once the GenAI system is in the field, organizations should compare the in-use rejection rate with the rejection rate observed during T&E. Dramatically different rates may indicate that human oversight isn't working properly (or, just as critically, it may reveal issues with system performance). For example, if you expect the system to produce incorrect answers 20% of the time but reviewers are flagging only 5% of the output, the disparity likely indicates automation bias, pressure to review outputs quickly, or one of the other factors that hinder human oversight.

Establish a quality control process. Oversight can also be enhanced by regularly assessing the quality of human reviews (are they identifying correct and incorrect outputs accurately?) and looking for evidence of automation bias (are reviewers actually assessing the outputs?). One quality

control technique is to introduce intentional errors every so often. If a reviewer fails to flag the response as incorrect, they're alerted that this was an error and they failed to catch the error. These periodic nudges are often enough to ensure that reviewers will carefully evaluate outputs and that automation bias doesn't take over. One caveat: organizations need to strike a careful balance, using these test cases in just the right measure. Too many, and the GenAI system's efficiency gains will suffer. Too few, and the process may have little impact on reviewer performance.

Build review time into the business case. Company leaders—and, ultimately, reviewers—often come to see oversight as a drain on a GenAI solution's value potential. Evaluations take time, they slow down the works, and they keep organizations from realizing all the gains they anticipated. By factoring review time into the solution's business case, companies set more realistic expectations for value. This makes it less likely that oversight gets thrown under the bus. And if the value potential is lower as a result, that's okay, too, as the richer cost-benefit analysis helps companies better prioritize solutions to develop. An up-front analysis also lets leaders know early on that a business case no longer closes. That way, they can cut bait before making an investment they won't recoup.

Take a risk-differentiated approach. Human oversight is time well spent—usually. An all-in, always-on, leave-no-output-unturned approach can cancel out all efficiency gains that a GenAI system can bring. But by designing oversight in a risk-differentiated way, companies can strike the right balance between review and efficiency. Differentiation can take different forms. It might be based on the purpose of the system (some systems, for instance, may need more review than others) or, drilling down deeper, the specific decision at play, with more review required for higher-risk output.

> By designing oversight in a risk-differentiated way, companies can strike the right balance between review and efficiency.

Leverage GenAI for oversight. In areas like data management, GenAI is already proving to be its own enabler. Perhaps the same could be true for oversight. For example, systems designers could build in a self-assessment capability. In effect, the GenAI system critiques itself, providing an assessment of confidence in its answer (say, through a confidence score), with lower certainty triggering more human review. GenAI-based review systems offer the advantages of speed, scale, and immunity to disincentives (they don't worry about the consequences of pushing the "escalate" button). Organizations that think carefully about how to combine GenAI-

based and human reviewers may find themselves with the most effective, efficient kind of oversight.

> Organizations that think carefully about how to combine GenAI-based and human reviewers may find themselves with the most effective, most efficient kind of oversight.

Educate users. Robust human oversight is fueled by knowledge: not only evidence for and against the output but also an understanding of the strengths and limitations of GenAI technologies and the implications of a system's different risks. Educating users also means sharing results from the T&E phase and providing insight into when the system performs well and when there tend to be gaps. And it means articulating—clearly and precisely—the purpose of a GenAI system or use case, so reviewers can identify not only inaccurate results, but also deviations from the intended function. Organizations should ensure that each GenAI system has a system card—documentation summarizing capabilities, intended use, limitations, risks, and T&E results—and make it easily accessible to all users.

Human oversight helps keep GenAI's value coming and its perils at bay. But it only works when it is carefully designed, not casually delegated. Companies that get oversight right make it an integral component of both GenAI systems design and risk mitigation. They foster vigilance where and when it matters most. And they empower reviewers to say something when they see something. GenAI systems aren't perfect; neither are humans. But with robust oversight, both technology and people can realize their potential—safely, reliably, and fully.

Authors

Steven Mills

Managing Director & Partner, Chief AI Ethics Officer, Global Leader, BCG Center for Digital Government
Washington, DC

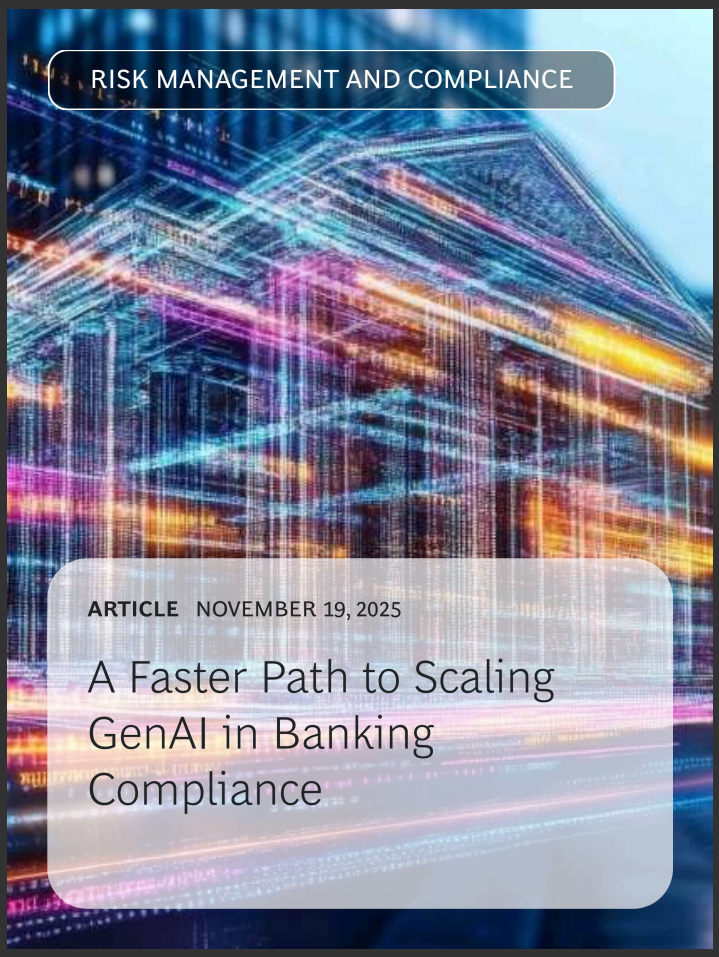
Noah Broestl

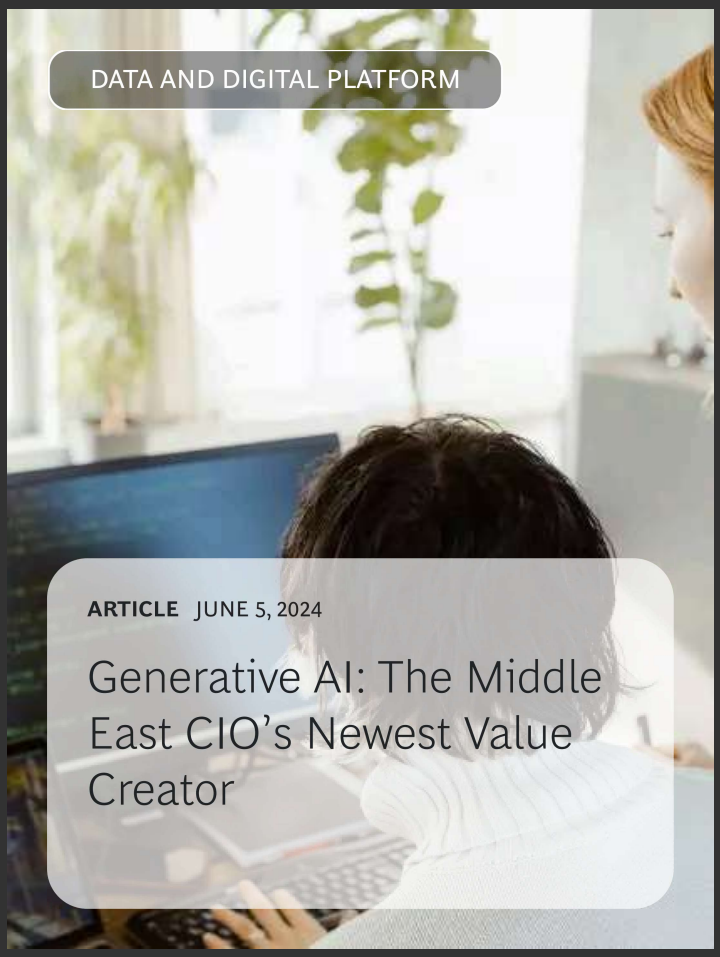
Partner and Associate Director, Responsible AI
Brooklyn

Anne Kleppe

Managing Director & Partner
Berlin

Related Content





ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963.

Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact. Our diverse global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.

Copy URL

Article 14: Human Oversight

Date of entry into force: **2 August 2026**

According to: **Article 113**

See here for a full implementation timeline.

SUMMARY +

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

2. Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this
Section.

3. The oversight measures shall be commensurate with the risks, level of autonomy and context of use of the high-risk AI system, and shall be ensured through either one or both of the following types of measures:

> (a) measures identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;
>
> (b) measures identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the deployer.

4. For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural

persons to whom human oversight is assigned are enabled, as appropriate and proportionate:

(a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;

(b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;

(c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;

(d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;

(e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

5. For high-risk AI systems referred to in point 1(a) of Annex III, the measures referred to in paragraph 3 of this Article shall be such as to ensure that, in addition, no action or decision is taken by the deployer on the basis of the identification resulting from the system unless that identification has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority. The requirement for a separate verification by at least two natural persons shall not apply to high-risk AI systems used for the purposes of law enforcement, migration, border control or asylum, where Union or national law considers the application of this requirement to be disproportionate.

← PREVIOUS NEXT →