SIDLEY

ARTIFICIAL INTELLIGENCE UPDATE

When "The Devil Made Me Do It" Is Not a Defense: Lessons in AI Governance and Organizational Oversight from an SDNY Decision

MAY 11, 2026

As companies increasingly integrate generative and agentic AI into core business functions, a May 7, 2026 decision from the United States District Court for the Southern District of New York[1] highlights several fundamental guardrails for corporate legal and compliance departments to consider. Although the case arose in the context of government decision-making, the opinion carries broader implications for any entity that embeds generative AI in its processes.

The opinion examined the process and output implications of the government's delegation of a sorting mechanism for grants that did or did not implicate DEI to ChatGPT in ways the court found raised constitutional issues. The court noted that the government "cannot escape liability … by scapegoating ChatGPT" because it was the government's "chosen instrument for purposes of this project," and the government's processes lacked sufficient human involvement, oversight, and validation.

The significance of the opinion does not turn on either the use of ChatGPT itself or the fact that it involved the government. Similar issues could arise from a company's use of any large language model (LLM), including internally developed models, or even industry-specific AI systems integrated into operational workflows. Companies deploying AI systems should consider whether they have adequate governance, oversight, and human involvement at all stages that are capable of withstanding scrutiny during litigation, enforcement actions, or regulatory review.

Case Overview

The case at issue, *American Council of Learned Societies v. National Endowment for the*

Humanities, involved allegations that the U.S. government unconstitutionally terminated more than 1,400 federal grants because they related to DEI. According to the opinion, U.S. government personnel used an LLM as part of the process for reviewing and identifying grants for potential termination, including by entering grant descriptions into the LLM with the following prompt: "Does the following relate at all to DEI? Respond factually in less than 120 characters. Begin with 'Yes.' or 'No.' followed by a brief explanation."

Plaintiff challenged the decisions to terminate grants based on the results of this LLM input. In granting summary judgment against the government, the court rejected the government's attempt to distance itself from responsibility for problematic generative AI outputs:

> The Government suggests that there is no real constitutional problem here because any viewpoint-based classification was ChatGPT's doing, rather than the Government's. That argument brings to mind, for someone of my generation, the great comedian Flip Wilson, whose character "Geraldine Jones" would excuse her behavior by saying, "The devil made me do it" That excuse did not work for Geraldine Jones, and it does not work for the Government.

The court's opinion reflects a basic, and broadly applicable, principle: A decision-maker cannot choose a tool to perform a task and then avoid responsibility by blaming the tool for the result. In reaching that conclusion, the court identified a number of deficiencies in the government's use of ChatGPT, with a focus on the human failures at issue. According to the opinion, the relevant personnel "did not examine any of the applications or underlying materials" associated with the grants, and there was "not a scintilla of evidence" that they undertook meaningful review of the AI-generated rationales before adopting and implementing them. The court also pointed to evidence that the individuals involved did not understand how ChatGPT itself interpreted concepts such as "DEI," and did not provide enough guidance to ChatGPT on the front end (i.e., through thoughtful prompting) to avoid obvious errors.

In the court's view, the issue was that AI-generated classifications and rationales became embedded into an operational decision-making process without sufficient development of prompts, review, contextual analysis, or understanding of how the outputs were generated. Put differently, there simply was not enough involvement from a "human in the loop" at each stage of the process.

Finally, the court underscored the extent to which the AI-assisted workflow itself became part of the evidentiary record. The opinion relied extensively on prompts and outputs in reconstructing how the underlying decisions were made and whether the outputs could be considered rational as a result. Notably, the court observed that certain materials, including those reflecting the use of ChatGPT, were not originally produced and that, when later produced, they confirmed that "misrepresentations … had been made to the Court" and that the government "had not

conducted a good-faith search of all relevant custodians and sources."

Practical Takeaways

1. **Establish clear governance structures for AI-assisted workflows involving human review, clear communication, and accountability.**

a. Prompt design itself may become a critical control point.

The opinion highlights that prompt design itself could be a governance issue, particularly where personnel lacking sufficient subject-matter expertise formulate prompts that invite outputs without considering relevant context. This becomes particularly significant where prompt design itself may materially influence the outcome. A failure to consider the nuances of a particular prompt, particularly without a robust review control, may introduce hallucinations or errors in ways that are less obvious than fabricated citations or objectively false factual assertions.

For example, here, the court's concern involved prompts asking the AI system to interpret and characterize a concept that could itself be subject to multiple interpretations and contextual considerations. The opinion specifically emphasized that the government personnel using ChatGPT did not define "DEI" for the model and did not understand how the model itself interpreted the term. The court treated that failure as significant because the undefined concept became the operative classification criterion for downstream decision-making. In this case, the prompts used caused the LLM to make what the court found to be serious qualitative and quantitative errors. As just one example, the LLM output categorized a study on whaling as being related to DEI. And the LLM found that an overwhelming number of existing grants met the criteria for termination.

The case therefore highlights the importance of implementing controls around who is authorized to formulate prompts for workflows that are part of organizational decision-making and how those prompts are considered or reviewed. Depending on the tool and use case, this may require personnel with sufficient subject-matter expertise, as well as familiarity with and training considering how prompts themselves may affect outputs, to ensure that relevant context and considerations are appropriately incorporated into the process.

b. Meaningful human review must involve more than nominal oversight.

The opinion also repeatedly noted the apparent absence of sufficient human review, validation, contextual analysis, and understanding of how AI-generated outputs were produced before those outputs became embedded into the decision-making process. A red flag noted by the court was the absence of a single example where a human reviewed an AI-generated rationale,

disagreed with the assessment, and chose a different course of action. The court's analysis also suggests that nominal "human involvement" alone may provide only limited protection if the review process is superficial or incapable of identifying unsupported or erroneous outputs.

Organizations using AI tools in operational workflows therefore should consider implementing and documenting meaningful human review procedures, including the following:

- approval and sign-off processes
- validation and testing protocols
- escalation requirements
- audit and monitoring mechanisms
- clear accountability structures and experience requirements for reviewing and operationalizing AI-generated outputs.

c. AI governance also requires clear communication and accountability structures.

Organizations should not assume that the use of generative or agentic AI meaningfully distances the organization from responsibility for resulting conduct. Here, the court placed responsibility for problematic outputs on the humans using the tool, not on the generative AI tool itself. The court also highlighted the disconnect between the individuals generating AI-assisted analyses and the individuals ultimately relying on those analyses in operational decision-making.

For each AI-assisted workflow, companies should establish clear lines of communication and accountability regarding:

- who is authorized to use AI tools
- who is responsible for prompt design
- who must review and validate outputs
- what information regarding the AI-generated process must be communicated to decision-makers
- when escalation or additional review is required
- which personnel ultimately retain responsibility for operational decisions informed by AI-assisted analyses.

Similar frameworks already exist across broader legal and corporate governance frameworks. For example, the Model Rules of Professional Conduct require legal knowledge, skill, thoroughness, and preparation reasonably necessary for representation, while fiduciary and duty-of-care principles require informed and reasonably diligent decision-making. These frameworks can be difficult to reconcile with an "AI made me do it" defense in a variety of contexts.

2. **Treat AI-assisted workflows as potentially discoverable from the outset.**

The court in *National Endowment* found generative AI prompts and outputs to be discoverable and relied on them extensively in evaluating how the underlying decisions had been made. Organizations should consider this as they increasingly deploy generative and agentic AI tools across many aspects of their businesses, including in sensitive areas such as investigations, compliance, HR, audit, claims review, vendor screening, and risk management. Potentially discoverable materials are being generated at an accelerated pace, at times without sufficient human validation. Those records later may become evidence in litigation or investigations, regardless of whether they are accurate, reliable, or ultimately adopted by the organization. And courts increasingly are requiring disclosure of generative AI use, whether in court filings or as part of discovery.

Organizations should understand where AI is being used, particularly in workflows involving sensitive or high-risk functions, and consider whether existing policies, employee training, validation procedures, monitoring mechanisms, retention practices, and litigation hold processes adequately account for AI-assisted business processes and AI-generated materials.

---

[1] *Am. Council of Learned Soc'ys v. Nat'l Endowment for the Humanities* , 2026 WL 1256545 (S.D.N.Y. May 7, 2026).

Contacts

If you have any questions regarding this Sidley Update, please contact the Sidley lawyer with whom you usually work, or

PARTNER

David A. Gordon

dgordon@sidley.com

Chicago +1 312 853 7159

PARTNER

Takayuki Ono

tono@sidley.com

**Not a registered foreign lawyer in Japan.*

Chicago +1 312 853 7296
Tokyo +81 3 3218 5096

COUNSEL

Matt S. Jackson

matthew.jackson@sidley.com

Chicago +1 312 853 4101

COUNSEL

Daniel Lim

daniel.lim@sidley.com

Washington, D.C.
+1 202 736 8699

ASSOCIATE

Kseniya K. Belysheva

kbelysheva@sidley.com

Los Angeles +1 213 896 6028