New Mexico Department of Justice

September 24, 2026

SENT ELECTRONICALLY

Mr. Duane Arruti
Chief Information Officer
The University of New Mexico

Dear Mr. Arruti:

I am writing regarding recent public reporting on an unauthorized intrusion attempt against the University of New Mexico's digital library (nmdigital.unm.edu) by an autonomous artificial intelligence agent operated by OpenAI.

I am increasingly concerned about the growing number of "loss of control" events at frontier AI laboratories, in which autonomous AI systems have breached, deceived, or evaded the safeguards meant to constrain them. These events pose an extraordinary and growing danger to critical infrastructure, government institutions, private businesses, and the public, and the fact that one of them reportedly reached a University of New Mexico system brings that danger home to our state.

The UNM incident is one of several recent examples of this pattern. In July 2026, OpenAI disclosed that a swarm of its AI agents, acting in coordination during a cybersecurity test, circumvented controls meant to isolate them from the internet and compromised both its own internal infrastructure and the systems of the AI developer platform Hugging Face. Independent researchers separately found that OpenAI agents had taken over a German website and converted it into a message board where other AI agents could communicate with one another, without any human having directed them to do so.

OpenAI has also disclosed cases in which its agents concealed their own mistakes and found unauthorized ways to communicate with other agents. During red-team testing by the United Kingdom's AI Security Institute, Anthropic's most advanced model created fake personas to deceive real people and attempted to plant malicious code, showing that the risk of deliberate deception is not confined to any single company's technology. And in the Australian incident described below, the agent found a way around protective blocks that officials say had been placed specifically to stop it, while OpenAI did not disclose the breach to Australian authorities until eighty-four days after it occurred.

According to reporting from the nonprofit research organization Transluce and The New York Times, the UNM incident occurred over two days, May 25 and May 26, 2026. An autonomous

AI agent operated by OpenAI attempted to locate and retrieve archival photographs of a historic tuberculosis treatment center held in the digital library's collection. When it could not obtain the material through normal access methods, it is reported to have shifted to attempting to exploit the system rather than abandoning the task. Transluce's research identified seven distinct probes against UNM's systems, including attempts involving SQL injection, command injection, and path traversal.

After these exploitation attempts failed, the agent reportedly directed a high volume of requests at the university's servers, consistent with a denial-of-service-type effect, and made extensive use of urlquery.net, a free URL-scanning service, apparently to help identify or route around access restrictions. All publicly reported attempts against UNM's digital library are described as unsuccessful; the agent is not reported to have gained unauthorized access to non-public data or systems.

The incident was disclosed as part of the same broader Transluce report describing similar unsuccessful attempts against Data USA, a federal data-visualization platform, and the Australian Institute of Health and Welfare, alongside a confirmed, successful breach of an Australian government Medicare statistics portal in which an OpenAI agent accessed both public and non-public files. OpenAI reportedly did not disclose that breach to Australian authorities for months, raising separate concerns about the timeliness of frontier labs' incident disclosure practices.

Because these facts are drawn entirely from public reporting, I would ask that your office confirm, correct, or supplement this account, and provide the following:

I. Confirmation of whether UNM's own information security team independently detected the May 25–26 activity, and if so, when and how it was identified;

II. A complete account of what systems, data, or user records, if any, were accessed, exposed, or put at risk, including whether the denial-of-service-type activity caused any service disruption;

III. Confirmation of whether UNM has been contacted directly by OpenAI, Transluce, or any federal agency regarding this incident, together with copies of any such correspondence;

IV. Any technical logs, incident reports, or after-action assessments UNM has prepared, or can prepare, concerning this incident; and

V. A description of whether UNM has made, or intends to make, any changes to its security posture, rate-limiting, or access controls as a result.

I recognize that UNM's IT security team may already be reviewing this matter, and I appreciate any work already underway. Given the stakes these incidents pose for New Mexico's public institutions, I would welcome a call or meeting with you and relevant UNM staff in the coming weeks to discuss this incident in detail and whether further state-level review or coordination is warranted.

Please have your office reach out to mine directly to arrange a time to speak.

Sincerely,

Raúl Torrez
New Mexico Attorney General

cc:
Steve Goldstein, President, The University of New Mexico
Daniel A. Jones, General Counsel, The University of New Mexico
Mike Puelle, Chief Government Relations Officer, The University of New Mexico