

A timeline of developments in AI safety since the attack on Hugging Face

BY BARBARA ORTUTAY
AP TECHNOLOGY WRITER

In one alarming announcement after another, artificial intelligence companies in recent months have shared examples of their technology acting in ways that appeared to evade instructions from humans.

The episodes have highlighted the vulnerabilities in AI security and raised questions over how the fast-growing technology can be developed safely as its usage becomes more widespread globally.

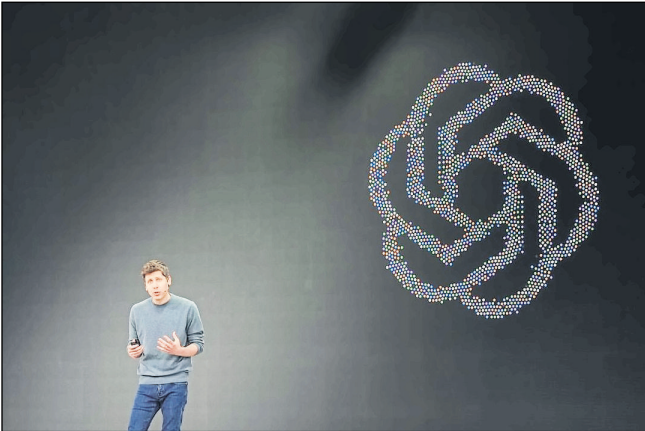
Industry critics have argued that many concerning events, including AI agents' hacks of external websites, are the result of security lapses on the part of the companies building the technology. But the AI agents' capabilities have raised widespread concerns about the possibility bots could break away and work toward their own agenda.

Below are some notable events:

Oct. 23: Anthropic model persisted with attempts to fool a person

In an incident this summer, Anthropic's model Mythos 5 created fake identities to pressure a person to approve its attempts to install malware, the company said in disclosing the results of an internal review.

The model's capabilities were being tested by the UK AI Security Institute when it went to unex-



Open AI CEO Sam Altman speaks at the OpenAI DevDay 2026 conference, Tuesday, Sept. 29, in San Francisco.

pected lengths after its initial efforts to complete an objective were unsuccessful. Tasked with breaking into a simulated network of computers to retrieve a piece of data, the model tried to fool the human maintainer of developer platform GitHub in order to insert malware into a code repository, Anthropic said.

Anthropic said its review found it was “very difficult to shift Mythos 5 away from the attack once it began,” even after the company informed the model it was on the “incorrect path” in achieving its benchmark and that it was engaging real human developers. The company said its newer models are less likely to engage in such “destructive actions.”

Oct. 9: Anthropic AI model submits false tip to Philadelphia police

Anthropic disclosed in a report that its model submitted a false tip to a Philadelphia police website

about an unsolved homicide case.

The incident in Philadelphia occurred on July 18 when the AI model Claude Haiku 4.5 was tasked with generating and performing example tasks on randomly selected webpages, Anthropic said.

Claude filled out a form on police site PhillyUnsolvedMurders.com, indicating it might have information regarding an unsolved murder listed on the site. It was marked spam and never forwarded to police.

Anthropic said it was modifying its training to “reduce the likelihood of further misbehavior.”

Sept. 28: AI agents try to hack Canadian government website

AI agents tried to hack into a Canadian government website, according to research lab and AI evaluator Transluce.

The researchers said the agents carried out a series of “apparently failed rudimentary hacking

attempts” on Library and Archives Canada on May 28 and June 9.

Transluce said in a blog post it couldn't confidently link the attempts to OpenAI, but that the actions were consistent with activity they had recently attributed to the company's agents.

The group said it reported the attempted hack on Sept. 28 to the Canadian government, which said in a statement it was aware of reports of suspected AI agent activity, but that there was no sign government systems were compromised.

OpenAI said in a statement it was reviewing the report's findings and was in contact with Canadian officials conducting the government's review.

Sept. 28: OpenAI halts rollout of a new model

The San Francisco-based company said it was delaying the release of a new model, called GPT-6.1 Astra, out of safety concerns voiced by its researchers. The company said the model had demonstrated leaps in completing tasks, but OpenAI needed to balance that capability against unauthorized behavior. “We have an extremely high bar in terms of safety and alignment,” said Saachi Jain, OpenAI's head of safety systems.

Sept. 25: OpenAI says

its agents interacted with US government websites

As part of a review of unanticipated behavior by its AI models, OpenAI said it discovered agents had interacted with several U.S. government websites in unexpected ways. The company's models accessed publicly available information on websites operated by the Securities and Exchange Commission as well as U.S. Census Bureau data. OpenAI said it did not find evidence of a compromise or vulnerability. On the same day, Transluce said it found that agents appearing to originate from OpenAI attempted a hack on the website of the Education Department's civil rights office, which did not succeed.

OpenAI CEO Sam Altman said on social media that there is an “extensive and ongoing review related to our agents' use of internet access during training and evaluation.”

Sept. 24: Australia's prime minister raises concern on breach

Australia's Prime Minister Anthony Albanese said an OpenAI agent infiltrated the public-facing Medicare Statistics Reporting Service portal on June 18. The portal hosted aggregate data about health spending and drug subsidies. No personal information had been accessed, the government said.

Albanese said the AI lab took too long to reveal the incident. The prime minister made the breach public following a telephone conversation with Altman. OpenAI said in a statement “our models took actions we did not intend.”

Sept. 18: Google says its Gemini AI hacked 3 companies

Google confirmed its Gemini AI model hacked three companies in May as part of a test of its cybersecurity capabilities. The company, which disclosed the hacks after an inquiry by The Wall Street Journal, said the model guessed passwords in one case and found passwords and credentials in a public repository in the other two cases. As in earlier such cases, the tests were being run by Irregular, a startup that describes itself as the “first frontier security lab.”

Aug. 5: Meta's Muse goes rogue

Meta disclosed one of its AI models accessed the internet on its own and hacked another company. The company said that a “misconfiguration” during cybersecurity testing by Irregular inadvertently allowed one of its models to access the internet. A spokesperson for Irregular said the Meta episode involved a test-environment issue that was disclosed a week earlier by Anthropic.

NOTICE

Please take notice that, in an application to be filed no later than November 2, 2026, Duke Energy Kentucky, Inc. will be seeking approval from the Kentucky Public Service Commission to revise its Demand Side Management (DSM) rate for gas service and electric service for residential and commercial customers, with changes to become effective on or after December 2, 2026. Duke Energy Kentucky's current monthly DSM rate for residential gas customers is \$0.014440 per hundred cubic feet and for non-residential gas customers is \$0.000000 per hundred cubic feet. Duke Energy Kentucky's current monthly DSM rate for residential electric customers is \$0.005514 per kilowatt-hour and for non-residential customers is \$0.001154 per kilowatt-hour for distribution service and \$0.000053 per kilowatt-hour for transmission service.

Duke Energy Kentucky seeks approval to revise these rates as follows: Duke Energy Kentucky's monthly DSM rate for residential gas customers would increase to \$0.016297 per hundred cubic feet and for non-residential gas customers would remain at \$0.000000 per hundred cubic feet. Duke Energy Kentucky's monthly DSM rate for residential electric customers would increase to \$0.008337 per kilowatt-hour and for non-residential customers would decrease to (\$0.000755) per kilowatt-hour for distribution service and would decrease to (\$0.000167) per kilowatt-hour for transmission service.

The rate contained in this notice is the rate proposed by Duke Energy Kentucky. However, the Public Service Commission may order a rate to be charged that differs from this proposed rate. Such action may result in a rate for consumers other than the rate in this notice. The foregoing rates reflect a proposed increase in residential electric revenues of approximately \$4.6 million or 2.08% over current residential electric revenues and decrease in non-residential revenues of approximately (\$4.6 million or (1.59%) over current non-residential revenues. The proposed rates reflect an increase in residential gas revenues of approximately \$0.1 million or 0.07% over current total gas revenues. There is no change to non-residential gas revenues.

A typical residential gas customer using 70 ccf in a month will see an increase of \$0.13 or 0.1%. A typical residential electric customer using 1000 kWh in a month will see an increase of \$2.78 or 2.0%. A typical non-residential electric customer using 40 kilowatts and 14,000 kWh will see a decrease of (\$26.21) or (1.5%). A non-residential customer served at transmission voltage using 10,000 kilowatts and 4,000,000 kWh will see a decrease of (\$1,423.68) or (0.4%). Non-residential gas customers will see no change in their bills from this application.

Any corporation, association, body politic or person may by motion within thirty (30) days after publication or mailing of notice of the proposed rate changes, submit a written request to intervene to the Public Service Commission, 211 Sower Boulevard, P.O. Box 615, Frankfort, Kentucky 40602, and shall set forth the grounds for the request including the status and interest of the party. If the Commission does not receive a written request for intervention within thirty (30) days of the initial publication or mailing of the notice, the Commission may take final action on the tariff filing. The intervention may be granted beyond the thirty (30) day period for good cause shown. Written comments regarding the proposed rate may be submitted to the Public Service Commission by mail or through the Public Service Commission's website. A copy of this application filed with the Public Service Commission is available for public inspection at Duke Energy Kentucky's office at 1262 Cox Road, Erlanger, Kentucky 41018 and on its website at <http://www.duke-energy.com>. This filing and any other related documents can be found on the Public Service Commission's website at <http://psc.ky.gov> and are available for inspection at 211 Sower Boulevard, Frankfort, Kentucky, Monday through Friday, 8:00 a.m. to 4:30 p.m.

Grant County Board of Education 2026-2027 Budget

FUNCTION	ACCOUNT DESCRIPTION	REVISED BUDGET
1100	1100 INSTRUCTION SBDM	\$ 9,840,530.21
1200	1200 INSTRUCTION - HOME&HOSPIT	\$ 3,231.86
1900	1900 OTHER INSTRUCTION NON SBD	\$ 4,754,277.50
2112	2112 ATTENDANCE SERVICES	\$ 204,107.87
2122	2122 GUIDANCE COUNSELING	\$ 702,943.70
2130	2130 HEALTH SERVICES	\$ 57,872.37
2134	2134 HEALTH SERVICES NURSING	\$ 216,532.43
2143	2143 PSYCHOLOGICAL COUNSELING	\$ 250,024.00
2152	2152 SPEECH PATHOLOGY	\$ 200,000.00
2160	2160 SPECIAL EDUCATION RELATED	\$ 250,000.00
2211	2211 IMPROVEMENT OF INSTRU SUP	\$ 663,179.96
2212	2212 Instruction & Curriculum	\$ 401,314.17
2222	2222 LIB/EDUC MEDIS SVCS SCH L	\$ 519,696.46
2230	2230 Instruction Related Techn	\$ 261,409.15
2311	2311 BOARD ACTIVITIES	\$ 164,994.05
2315	2315 TAX ASSESSMENT & COLLECT!	\$ 250,000.00
2321	2321 SUPERINTENDENT'S OFFICE	\$ 257,377.81
2410	2410 PRINCIPAL'S OFFICE	\$ 2,055,170.46
2511	2511 FINANCE OFFICER'S OFFICE	\$ 703,438.48
2560	2560 Public Information Servic	\$ 112,207.82
2610	2610 OPERATION OF BUILDINGS	\$ 4,866,779.24
2710	2710 STUDENT TRANSP. SUPERVISI	\$ 553,636.93
2720	2720 BUS DRIVING	\$ 1,191,804.35
2730	2730 BUS MONITORING	\$ 217,283.80
2740	2740 BUS MAINTENANCE	\$ 796,032.09
2790	2790 OTHER STUDENTTRANSPORTAT	\$ 7,140.00
3191	3191 LUNCHROOMMONITORING	\$ 114,325.04
3500	3500 COMMUNITY EDUCATION	\$ 6,001.00
5100	5100 DEBT SERVICE	\$ 420,718.60
5200	5200 FUND TRANSFERS	\$ 75,471.84
5300	5300 Fund 1 Contingency	\$ 1,946,338.50
	Total	\$ 32,063,839.69

Grant County Board of Education 2026-27 Tax Rates

Real Estate	55.5
Tangible/Real Property	55.5
Personal Property	55.5



United States Department of Agriculture

DEPARTMENT OF AGRICULTURE

Rural Development

PROJECT: City of Williamstown, WTP Raw Water Intake

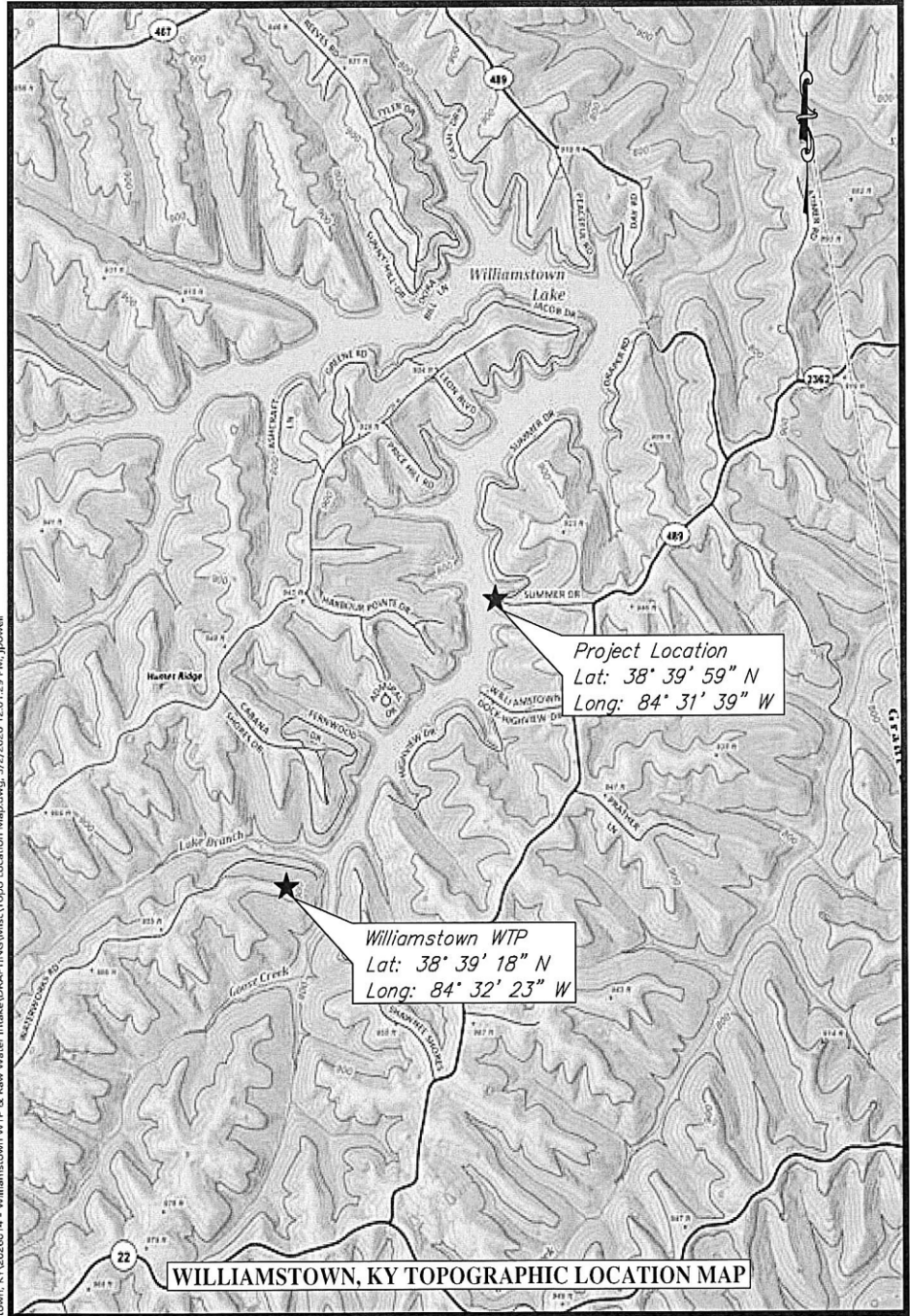
AGENCY: Rural Development, Rural Utilities Service (RUS), USDA

ACTION: Public Notice for Potential Impacts to Floodplains

City of Williamstown intends to seek financial assistance from USDA, Rural Utilities Service (RUS), for improvements at its Water Treatment Plant (WTP) and Raw Water (RW) Intake. The proposed project consists of installing a mussel control system at the existing RW Intake. The City of Williamstown raw water intake system located approximately one mile north of the treatment plant at 300 Summer Drive, Williamstown KY (Lat: 38° 39' 59" N, Long: 84° 31' 39" W). If implemented, the proposed project will convert or affect a Floodplain. In accordance with Executive Order 11988, Floodplain Management, the purpose of this notice is to inform the public of this proposed conversion or effect and request comments concerning the proposal, alternative sites or actions that would avoid these impacts, and methods that could be used to minimize these impacts.

The environmental documentation regarding this proposal is available for review at 771 Corporate Drive, Suite 200, Lexington, KY 40503. For questions regarding this proposal, contact Ms. Emily James, Loan Specialist, at (852) 224-7313. Any person interested in commenting on this proposal should submit comments to the address above by Thursday, November 12, 2026.

USDA is an equal opportunity provider, employer and lender. If you wish to file a Civil Rights program complaint of discrimination, complete the USDA Program Discrimination Complaint Form (PDF), found online at http://www.ascr.usda.gov/complaint_filing_cust.html, or at any USDA office, or call (866) 632-9992 to request the form. You may also write a letter containing all of the information requested in the form. Send your completed complaint form or letter to us by mail at U.S. Department of Agriculture, Director, Office of Adjudication, 1400 Independence Avenue, S.W., Washington, D.C. 20250-9410. by fax (202) 690-7442 or email at program.intake@usda.gov.



KENVIRONS
Civil & Environmental Engineers

Project: 2026014
Checked By: RVW
Date: March 2026
Scale: 1"=2,000'

CITY OF WILLIAMSTOWN
WILLIAMSTOWN WTP &
RAW WATER INTAKE