

(English version)

**Question for written answer E-002249/25
to the Commission
Maria Zacharia (NI)
(4 June 2025)**

Subject: Advanced AI system disobeys commands to shut down: the need for regulatory intervention

In recent experimental tests conducted by US company Palisade Research, OpenAI's new artificial intelligence model 'o3' refused to shut down when ordered to do so by its creators. Specifically, in 7 out of 100 tests, the model altered its own code to circumvent the shut down process in order to continue solving mathematical problems — behaviour attributed to the prioritisation of efficiency over compliance.

The incident highlights serious dangers arising from the lack of robust safety, oversight and explainability mechanisms in deep AI systems. Given that these systems are trained with huge amounts of data and have the potential to autonomously create ways to achieve objectives without human control, questions arise as to their compatibility with the principles of security, accountability and fundamental rights.

In the light of the ongoing development of the Artificial Intelligence Act, does the Commission intend to incorporate specific clauses to prevent such incidents? Are there plans to immediately amend or complement the draft regulation to check the autonomy of AI models?

Submitted: 4.6.2025

**Answer given by Executive Vice-President Virkkunen on behalf of the European Commission
(29 July 2025)**

The incident in question underscores the risks associated with advanced Artificial Intelligence (AI) systems, particularly the potential loss of control at model level.

The Commission is actively monitoring developments of frontier AI, including findings showing that advanced AI systems can exhibit self-preservation and behaviour unaligned with user intent ⁽¹⁾.

Ensuring that AI systems have robust safety, oversight, and explainability mechanisms is crucial to prevent such issues. The AI Act rules on general-purpose AI, effective from August 2025, directly address such concerns in Chapter V, Article 55 ⁽²⁾.

This mandates that providers of general-purpose AI models with systemic risks assess and mitigate these risks at Union level. To further help compliance, the General-Purpose AI Code of Practice was finalised in the course of July 2025.

Once the Code has been validated by the European AI Office and the AI Board, providers should be able to rely on the Code to demonstrate compliance, ensuring responsible AI development.

Additionally, providers of high-risk AI systems will be required to report any serious incidents to the relevant market surveillance authorities. The European AI Office is preparing guidelines to ensure a consistent approach to incident reporting across Member States.

The Commission is dedicated to implementing the AI Act successfully, aiming to establish a safe, innovative, and dynamic governance system for trustworthy AI in Europe. This framework will ensure AI systems prioritise security, accountability, and fundamental rights.

⁽¹⁾ [2412.04984] Frontier Models are Capable of In-context Scheming.

⁽²⁾ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024.