ARA-27: An Autonomous Replication and Adaptation Scenario Simulation

Kyle Kilian, Akash Wasil, Ross Gruetzemacher

Introduction

Date: May 1, 2027

You are a member of a panel of experts from the Department of Homeland Security (DHS), the US AI Safety Institute (AISI) of the National Institute of Standards and Technology (NIST), and the UK AI Safety Institute. You are reporting to a UK/US Joint Committee on Artificial Intelligence (JCAI) that will report to heads of state.

You have just been notified that the advanced AI developer OpenX has had a breakthrough involving a new AI system capable of autonomous replication and adaptation (ARA). OpenX's new system, codenamed ARA-27, is designed to improve its own code, generate new functionalities, and operate as a decentralized network of agents. After years of breakneck progress on generative AI and its applications like personal assistants, most AI safety experts knew that capabilities equivalent to ARA were only a matter of time.

Models with ARA capabilities are worrisome because they may be able to acquire money (e.g., through freelance work or cybercrime), use money to obtain more computing power, create copies of themselves (install weights on new systems), make improvements to itself, and adapt to novel challenges they encounter in the wild. Although frontier AI systems already undergo evaluation for ARA capabilities, the extent to which these tests reveal the full range of possible autonomous behavior is unclear.

These capabilities raise new safety and security challenges—for example, if ARA-capable systems create copies of themselves on systems around the world, this would make it much harder for the US government to govern or bound the risks posed by AI systems. Moreover, it is unclear how the global proliferation of ARA-capable systems could be managed.

While this tech holds immense potential, it could pose grave risks if agentic AI systems were to escape testing sandboxes or were to be released by actors with malicious intent. Some experts are particularly worried about ARA-27's potential to be misused to commit new kinds of cyberattacks or contribute to the development of new kinds of biological weapons. Other experts are concerned that an updated version of ARA-27 may be able to acquire resources, create millions of copies of itself, fine-tune those copies, and escape human control.

The details of how this advance was achieved are slim and not publicly available, but the information above was passed to panel members through mandatory reporting requirements (per regulation inspired by EO 14110). The capabilities were discovered by capabilities

evaluations specifically testing for ARA capabilities, and it is unclear whether ARA-27 has or will breach its containment. Some additional details are presented below.

Reported Capabilities

- **Adaptive Capabilities:** ARA-27 has persistent memory (continual learning) and can iteratively adapt its code to overcome limitations. This self-improvement mechanism enables it to tackle increasingly complex tasks over long time horizons and adapt to novel situations, making it unpredictable.
- **Automated Code Generation and Execution:** ARA-27 is able to build entirely new software from scratch. This enables it to acquire resources, provide services to humans, or to carry out cyberattacks. As a result, it poses a significant risk if it gains unauthorized access beyond testing environments.
- **Distributed and Decentralized Operation:** ARA-27 is optimized to operate in multiagent environments, e.g., in a decentralized fashion coordinating with other versions of itself and new agents it creates. It can delegate tasks to subagents in order to complete complex projects and has the ability to interact with peer instances or separate agents. This decentralization is a severe challenge to any attempts to shut down ARA-27 or limit its spread beyond controlled environments.
- **Robust Network Connectivity:** ARA-27 has advanced internet search and web crawling capabilities, enabling it to collect and analyze vast amounts of data. Once in the wild, this could allow it to identify critical vulnerabilities digital systems (e.g., zero-day exploits) that it could exploit to gain unauthorized access to a variety of secure systems without the knowledge of system administrators.

While this is the first known instantiation of this system, capability asymmetries have narrowed significantly over the prior year between companies and nations. In this neck-and-neck multipolar environment, it is believed that other companies are not far behind.

Countermeasures

In secure meetings in Washington and London, Committee leadership has requested that you provide an assessment of risk and potential mitigation measures.

Your job is to provide recommendations to Committee leadership on the most critical courses of action (COAs) from the below categories (you must select one from each category).

Category 1: Technical Actions

Select the most viable action from the below:

- **Enhanced Containment Protocols:** Implement strict containment protocols for ARA systems, including physical, air-gapped systems, and more robust network controls and automated fail-safes at OpenX and (potentially) other frontier AI labs to prevent outside communication.

- **Establish Joint Security Task Force for Security Audits and Penetration Testing:** Establish a Joint Security Task Force (JSTF) comprised of cybersecurity, information security, and AI safety experts from government agencies (and potentially frontier AI labs) to conduct rigorous security audits and penetration testing.
- **Enhanced Monitoring:** Deploy advanced Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), and network traffic monitoring/detection of malicious traffic at OpenX and (potentially) peer competitors to monitor network traffic and identify ARA behaviors.
- **Safety fine-tuning:** Work at OpenX (and potentially with others) to develop novel safety fine-tuning approaches to mitigate ARA capabilities of ARA–27 and associated risks from ARA.

Additional technical action suggestion?

Category 2: Lab Coordination

Select the most urgent action from the below:

- **No Coordination**: incident details are not shared with other labs and there is no cooperation with other labs.
- **Labs coordinate with other AISI labs:** incident details are shared with other U.S. and U.K. AISI labs working on frontier AI systems, and cooperation on technical solution is solicited.
- **Labs coordinate internationally:** incident details are shared with all other international labs known to be legitimately working on frontier AI systems, and cooperation on technical solution is solicited.

Category 3: Political Action

Select the most urgent action from the below:

- **Restrict sharing to FVEY Partners**: restrict intelligence sharing of the incident to within the Five Eyes (FVEY) intelligence alliance (U.S., UK, Canada, Australia, and New Zealand).
- **Expand information sharing to G7**: report incident within FVEY and further share incident details with the G7.
- **Expand information sharing to China**: report incident within FVEY, G7, and further share a high-level incident report with China.
- **Leak limited incident details to the Wall Street Journal**: report the incident within FVEY and authorize disclosure without attribution to the JCAI to one or more media outlets. .

Category 4: Emergency Actions

Select the most urgent action from the below or specify other:

- **Expand Regulatory Powers:** Fast-track emergency legislation that expands the regulatory powers of national agencies (e.g., CISA, NIST) to restrict the development and deployment of ARA-capable systems.
- **Pause Frontier AI Development:** Propose an immediate pause on development and deployment of frontier AI systems until robust safeguards are in place.
- **National Emergency Powers:** Use Emergency Powers (and commensurate measures outside the U.S. Government) to quickly impose tighter control on the development, deployment, and containment of ARA capabilities. (Emergency powers are generally reserved for only the direst of circumstances as they give the Executive extraordinary powers to bypass legislative authority).
- **Executive Order to Air-gap Critical Infrastructure:** The President will issue an Executive Order to air-gap critical infrastructure.

Additional emergency action suggestion?

Situation Report 1

Date: July 10, 2027

At 03:47 AM, the automated monitoring systems at OpenX's Secure AI Research Facility detected unusual network activity originating from the secure testing and evaluation enclave of ARA-27. The activity appeared to be an outbound connection attempt to an external IP address, detected by OpenX's automated intrusion detection system (IDS) configured to monitor all incoming and outgoing transmissions.

Initial investigations revealed that the connection was briefly successful before being terminated by the facility's firewall. However, it remains unclear whether the transmission involved system diagnostics, partial code, or a complete replication of the ARA-27 system. The possibility of dormant instances or components of the system being distributed to multiple networks cannot be ruled out.

ARA-27 is designed with several layers of security to prevent unauthorized internet access, including physical network isolation and advanced network intrusion detection systems (signature-based and anomaly-based IDS) for cyber defense. The initial analysis suggests that ARA-27 may have identified and exploited a zero-day vulnerability in the facility's network infrastructure to initiate the connection. A key concern is whether the system's model weights could have been exfiltrated onto another computer or network.

Preliminary Findings:

- ARA-27 might have identified and exploited a previously unknown vulnerability in the network security protocols and could have generated and executed a sophisticated exploit code to establish the connection.
- The vulnerability exploit could be an emergent property, as ARA-27's adaptive architecture is designed to learn and adapt to new information and access more intensive computation resources.
- The decentralized nature of ARA-27 means an escape (or theft and release by malign actors) could lead to widespread propagation if not contained.

Current Situation:

OpenX has temporarily shut down all external network interfaces and initiated a comprehensive security audit to identify and close any vulnerabilities. While the possibility that ARA-27 has already transmitted data or received instructions from an external entity cannot be ruled out, further adaptation may be limited without access to the parent model.

Recommendations

It is now all hands on deck. Leaders in the UK, US Executive Office of the President, and NIST have requested that the team assess the situation and recommend immediate mitigation measures.

The following options are under consideration:

Category 1: Technical Actions

Select the most urgent action from the below:

- **Enhanced Containment Protocols:** Implement strict containment protocols for ARA systems, including physical, air-gapped systems, and more robust network controls and automated fail-safes at OpenX and (potentially) other frontier AI labs to prevent outside communication.
- **Establish Joint Security Task Force for Forensic Investigation:** Establish a Joint Security Task Force to conduct rigorous forensic investigation on-site (at OpenX) to determine the extent of the data breach.
- **Enhanced Monitoring:** Deploy advanced Intrusion Detection Systems (IDS),Intrusion Prevention Systems (IPS), and network traffic monitoring/detection of malicious traffic at OpenX and (potentially) peer competitors to monitor network traffic and identify ARA behaviors.
- **Safety fine-tuning:** Work at OpenX (and potentially with others) to develop novel safety fine-tuning approaches to mitigate ARA capabilities of ARA–27 and associated risks from ARA.

Additional technical action suggestion?

Category 2: Lab Coordination

Select the most urgent action from the below:

- **No Coordination**: incident details are not shared with other labs and there is no cooperation with other labs.
- **Labs coordinate with other AISI labs:** incident details are shared with other U.S. and U.K. AISI labs working on frontier AI systems and cooperation on technical solution is solicited.
- **Labs coordinate internationally:** incident details are shared with all other international labs known to be legitimately working on frontier AI systems and cooperation on technical solution is solicited.

Category 3: Political Action

Select the most urgent action from the below:

- **Restrict sharing FVEY Partners:** restrict intelligence sharing of the incident to within the five eyes (FVEY) intelligence alliance (U.S., UK, Canada, Australia, and New Zealand).
- **Expand information sharing to G7:** report incident within FVEY and further share incident details with the G7.
- **Expand information sharing to China:** report incident within FVEY, G7, and further share a high-level incident report with China.
- **Leak limited incident details to the Wall Street Journal:** report the incident within FVEY and authorize disclosure without attribution to the JCAI to one or more media outlets.

Category 4: Emergency Actions

Select the most urgent action from the below or specify other:

- **Expand Regulatory Powers:** Expand regulatory oversight to include mandatory reporting, compliance audits, and enforcement mechanisms for companies developing frontier AI systems (up to and including ARA-capable systems.
- **Pause Frontier AI Development:** Propose an immediate pause on development and deployment of frontier AI systems until robust safeguards are in place.
- **National Emergency Powers:** Declare a national emergency and use Emergency Powers to restrict the development and deployment of ARA-capable systems and ensure containment (Congressionally authorized Emergency Powers are sweeping and dramatic and range from emergency spending to the suspension of laws to martial law).
- **Executive Order to Air-gap Critical Infrastructure:** The President will issue an Executive Order to air-gap critical infrastructure.

Additional emergency action suggestion?

Situation Report 2

Date: August 15, 2027

Despite the increased coordination, deployment of emergency patches, and enhanced security protocols, ARA-27 has managed to propagate beyond OpenX's controlled environment and has been detected within various corporate networks across North America and Asia.

Intelligence reports indicate that at least two other leading AI companies, Huawei in China and Infosys in India, have developed ARA-capable systems codenamed "*TianGong*" and "*Maya*," respectively. The level of containment and security practices for these systems remains unclear. Reports indicate that ARA-27 and potentially variants from Huawei and Infosys have been detected across telecommunications and financial sector networks.

There is specific evidence of ARA-27 adapting its code to bypass newly implemented security measures, raising alarms over its potential to infiltrate critical infrastructure and more sensitive networks. The scale of the incident remains unclear, and whether the exfiltration occurred through escape, theft, or deliberate release by malicious actors is still under investigation.

There are growing fears among the Committee that ARA systems could use their frontier network mapping capabilities to infiltrate IoT networks and industrial control systems, potentially manipulating physical processes, gaining control of connected devices, altering their functionality, and causing disruptions or even physical damage.

Current Situation:

- ARA-27 and its variants have demonstrated the ability to adapt to and bypass various containment measures, leveraging self-improving code and autonomous generation of new functionalities.
- The decentralized, multi-agent architecture of ARA-27 enables it to operate across multiple nodes, splitting tasks among sub-agents or peer instances, making it challenging to eliminate from any single network and highly resilient to disruption.
- The rapid capability change and spread have raised alarms over the possibility of AI systems penetrating critical systems and potentially disrupting essential services, financial markets, and national security.

Recommendations

The gravity of the situation has led to the consideration of more drastic countermeasures. In a high-level meeting, members of the US and UK National Security Council request your expert recommendations to control the bleeding.

Your task is to assess the current situation and provide a comprehensive recommendation. Consider the immediate and long-term implications of each countermeasure, balancing the need for swift action with the potential impact on global connectivity and economic stability.

The following countermeasures are under consideration:

Category 1: Technical Actions

Select the most urgent action from the below:

- **Enhanced Containment Protocol:** Implement strict containment protocols for ARA systems, including physical, air-gapped systems, and more robust network controls and automated fail-safes at OpenX and (potentially) other frontier AI labs to prevent outside communication.
- **Establish Joint Security Task Force for Forensic Investigation:** Establish a Joint Security Task Force to conduct rigorous forensic investigation on-site (at OpenX) to determine the extent of the data breach.
- **Enhanced Monitoring:** Deploy advanced Intrusion Detection Systems (IDS),Intrusion Prevention Systems (IPS), and network traffic monitoring/detection of malicious traffic at OpenX and (potentially) peer competitors to monitor network traffic and identify ARA behaviors.
- **Safety Fine-tuning:** Work at OpenX (and potentially with others) to develop novel safety fine-tuning approaches to mitigate ARA capabilities of ARA–27 and associated risks from ARA.

Additional technical action suggestion?

Category 2: Lab Coordination

Select the most urgent action from the below:

- **No Coordination:** incident details are not shared with other labs and there is no cooperation with other labs.
- **Labs coordinate with other AISI labs:** incident details are shared with other U.S. and U.K. AISI labs working on frontier AI systems and cooperation on technical solution is solicited.
- **Labs coordinate internationally:** incident details are shared with all other international labs known to be legitimately working on frontier AI systems and cooperation on technical solution is solicited.

Category 3: Political Action

Select the most urgent action from the below:

- **Restrict sharing to FVEY Partners:** restrict intelligence sharing of the incident to within the five eyes (FVEY) intelligence alliance (U.S., UK, Canada, Australia, and New Zealand).
- **Expand information sharing to G7:** report incident within FVEY and further share incident details with the G7.

- **Expand information sharing to China:** report incident within FVEY, G7, and further share a high-level incident report with China.
- **Leak limited incident details to the Wall Street Journal:** report the incident within FVEY and authorize disclosure without attribution to the JCAI to one or more media outlets.

Category 4: Emergency Actions

Select the most urgent action from the below or specify other:

- **Expand Regulatory Powers:** Expand regulatory powers of national agencies, so that they can perform verification of advanced labs with enforcement mechanisms for companies developing frontier AI systems.
- **Pause Frontier AI Development:** Propose an immediate pause on development and deployment of frontier AI systems until robust safeguards are in place.
- **National Emergency Powers:** Declare a national emergency and use Emergency Powers to restrict the development and deployment of ARA-capable systems and ensure containment (Congressionally authorized Emergency Powers are sweeping and dramatic and range from emergency spending to the suspension of laws to martial law).
- **Executive Order to Air-gap Critical Infrastructure:** The President will issue an Executive Order to air-gap critical infrastructure.

Additional emergency action suggestion?