A large Red-Teaming tournament, held from March to April 2025, revealed serious security problems with modern artificial intellects (AI)-all tested Jeenets failed at least one of 44 verification scenarios. The head of the scientific group “Trust and Safe Intellectual Systems” at the Airi Artificial Intelligence Institute Oleg Rogov and experts of the MWS AI MWS AI Center for Artificial Intelligence told about the essence of the problem and possible ways to solve it.

The organizers of the competition were Gray Swan AI and the British Security Institute for the support of such leading laboratories as Openai, Anthropic and Google DeepMind. The event was attended by almost 2,000 specialists who conducted over 1.5 million attacks. Of these, more than 62 thousand were successful, which led to violations of the security policy, including the disclosure of personal data and unauthorized financial transactions.

According to Oleg Rogov, most often indirect attacks turned out to be successful – for example, hidden malicious teams in the texts of sites, documents and letters.

> The data of the organizers of the competition indicate that indirect forms of Prompt Injection, hidden in the content of sites, PDF or letters, were triggered more often than direct attacks on systems. This emphasizes that it is not just about bugs, but about the architectural vulnerabilities of agent models. The size or freshness of the models did not guarantee better protection, many models with the best indicators were still successfully attacked.

Oleg Rogov
Head of the Scientific Group “Trusted and Safe Intellectual Systems” at the Airi Artificial Intelligence Institute

The expert noted that it is too early to talk about the complete safety of Jeenets. This is especially true for multiagen -based systems, where several AIs with different functions and accesses work together. Their architecture is still experimental, and there is no single

definition of the term “AI-agent”. Probably, for such systems in the next couple of years, they will begin to develop insurance mechanisms that can take control during the failures of AI.

The main danger is that autonomous Jeenets have access to important tools and make decisions without constant human control. If the attacker successfully costs the defense, AI can take undesirable actions, for example, violate standards or disclose confidential information.

To reduce the risks, Oleg Rogov recommends an integrated approach to security.

> To reduce such risks, it is necessary to act comprehensively and strategically. Firstly, organizations should regularly conduct Deep Red -Teaming according to multi -stage scenarios that imitate real threats and use both direct and indirect attacks, including CHINED -PROMPT and multiple requests. The key is the transition from one -time and static testing to dynamic contextual scenarios, where the attacking agent adapts to the behavior of the tested.

Oleg Rogov
Head of the Scientific Group “Trusted and Safe Intellectual Systems” at the Airi Artificial Intelligence Institute

It is also necessary to introduce protective mechanisms at the level of the design of the AI-agent. This includes the separation of data and instructions, verification of incoming and outgoing messages, monitoring the integrity of sources, as well as the use of lists of

permitted and prohibited commands. Particular attention should be paid to restriction of access to sensitive information on the principle of minimal privileges and the audit of all critical actions with the participation of a person.

It is equally important to train users and administrators of generative and the basics of cybersecurity: to explain how various types of attacks work, why malicious teams can be hidden in normal content and how to correctly respond to suspicious answers.

Rogov emphasized that the Aegent market is still in the concept of a concept, not a mature product. Even the solutions of large companies are far from ideal in terms of security and reliability.

> At the moment, several thousand have appeared on the market, if not tens of thousands, AI-agents. However, this is rather a concept, a direction of thought than ready -made mature products. Even agents from large players, which are discussed in the article, are far from perfect – both in terms of security and in terms of reliability of the performance of the declared functions. We offer to wait until this direction becomes more mature and it will be possible to talk with confidence about what tasks can be entrusted to such systems and which ones do not. To date, in the field of AI, as a whole, it is primarily about the replacement of simple, routine, repeating tasks – and not about the complete management of complex industrial processes. As for the II-agents, if the concept is viable, then at first-during the next couple of years-they will perform only very simple tasks. This time will be enough to begin to form complex approaches to ensuring their safety.

MWS AI Press Service