AI can learn unwanted things, even if the data for training seem safe, warns Oleg Rogov, candidate of physical and mathematical sciences and director of the Sail safe artificial intelligence of the Airi and Mtusi Institute. This is called “subconscious training” – when algorithms find hidden templates in the data that can lead to errors.

For example, the chat bout of the support service may suddenly start issuing toxic answers due to such hidden ties in training. Or the medical system can involuntarily ignore the symptoms in some patients, which is dangerous and can cause legal problems. It is especially difficult to notice such errors, since they are “hidden” inside the model itself, and not in obvious data.

> It is critically important to double-check the advice of non-core and systems, for example, publicly accessible chat bots to obtain answers in the field of medicine or finance with authoritative sources and expert opinion. For example, for a doctor, even a profile tool is always a second opinion. Services using unverified synthetic data should be avoided.

Oleg Rogov
Director of the Laboratory of Safe Artificial Intelligence “Sail” of the Institute Airi and MTUSI

The expert emphasizes that in order to increase safety and trust, AI must be made more transparent – to publish reports on tests, use “sandboxes” to check and tools showing how decisions are made. In critically important areas, it is worth applying the tested rules and a modular approach to training.

The AZR method is useful in education, in which AI is learned only on verified data-for example, mathematics or programming-to avoid dangerous errors due to harmful texts.