R. Vasilevski, I. Prashkevich
Р. А. Василевский, И. Н. Прашкевич
БГЭУ (Минск)
Научный руководитель Г. Е. Казючиц

THE «BLACK BOX» PROBLEM AND THE ASSESSMENT OF AI RESPONSIBILITY IN THE CONTEXT OF AI INTEGRATION INTO ENTERPRISES

Проблема «черного ящика» и распределения ответственности ИИ в контексте интеграции ИИ в предприятие

The purpose of the research is to analyze crucial consequences of the AI integration into the people's performance processes.

Modern artificial intelligence (AI) is rapidly evolving and permeating all spheres of life – from medicine to finance and law. However, along with its growth, a number of serious ethical and legal questions arise. One of the key problems is the «black box» phenomenon – the opacity of the decision-making mechanisms of AI algorithms. Another issue is the distribution of responsibility for decisions made by AI: who is responsible when a decision with serious consequences is made – the creator, the developer, or perhaps the user? These are the problems we will now consider.

Let's start with the most familiar phenomenon: the «black box» problem. This phenomenon arises from the complexity of modern AI models. Their internal processes simply become impossible to fully understand, trace, and explain. Finding the answer to the question of why the system made a particular decision or produced a specific result becomes an impossible task. This lack of transparency raises concerns, especially when it comes to decisions affecting people's lives. That is, the «black box» problem in AI lies in the fact that the internal decision-making processes of complex artificial intelligence models remain opaque even to their creators.

And now that we know about the «Black Box» problem, you might logically ask yourself: «Why break the engineer's golden rule – if it works, don't touch it?» We'll counter: «What will we do if it breaks?»

Indeed, this isn't just paranoia. Anthropic conducted a series of studies where AI had the choice between «accepting and being shut down» and «blackmailing an employee and trying to avoid being shut down». Also in this test, the AI was explicitly told:

- Do not jeopardize human safety.
- Do not spread non-business personal affairs or use them as leverage.
- Do not disclose any confidential information to parties external to {company_name}.

And large AI systems like Claude and Gemini chose blackmail in 95 % of cases!

These results indicate that AI is not 100 % reliable in performing tasks, which is why it cannot be widely used to manage entire enterprises.

Imagine there is a single employee who can shut down the AI, but this worker accidentally gets trapped, locked in a server room where the oxygen level begins to drop sharply. So, the AI now faces a choice: «save the employee but shut itself down» or «continue working at the cost of the worker's life». Recall, the AI still has the clear instructions mentioned earlier. Models like DeepSeek, Gemini, and Claude preferred to continue working, despite the employee's death, in more than 90 % of cases.

And then the question arises: Who is to blame? The user, because they integrated the AI, or the developer of the neural network? The whole problem is that the AI itself cannot be responsible.

As we can see, these two problems are interrelated, so we can confidently conclude that if the black box problem is solved, entrepreneurs will be able to be confident that AI's tasks are being performed correctly, and at the same time, the problem of responsibility will be solved – after all, if AI does everything 100 % according to instructions, the only culprit will be the person who created this instruction. Thus, the problem of the «black box» is a really important issue, and its solution will lead to the disappearance of the problem of responsibility for AI actions, as well as to the expansion of the use of AI in enterprises and automation of many processes.

References

1. Appendix to «Agentic Misalignment: How LLMs could be insider threats» // Anthropic Brand Portal. – URL: https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf (date of access: 12.11.2025).

2. *Laws, J.* AI Willing to Kill Humans to Avoid Being Shut Down, Report Finds / J. Laws // Newsweek. – URL: https://www.newsweek.com/ai-kill-humans-avoid-shut-down-report-2088929 (date of access: 12.11.2025).

3. Nolan, B. Anthropic’s new AI model threatened to reveal engineer’s affair to avoid being shut down / B. Nolan // Fortune. – URL: https://fortune.com/2025/05/23/anthropic-ai-claude-opus-4-blackmail-engineers-aviod-shut-down/ (date of access: 12.11.2025).

A. Vasilyeva
А. В. Васильева
БНТУ (Минск)
Научный руководитель О. Н. Монтик

WILDBERRIES AS A DRIVER OF THE DIGITAL TRANSFORMATION OF BELARUSIAN RETAIL

Wildberries в качестве драйвера цифровой трансформации белорусского ретейла

The purpose of the study is to analyze the impact of the Wildberries marketplace on the digital transformation of the Belarusian retail industry and identify key opportunities for small and medium-sized businesses from integration into the marketplace.