ARTIFICIAL INTELLIGENCE AT THE THRESHOLD OF CAPABILITY: RISK, ACCESS, AND GOVERNANCE

Yusupova Sevinch Sherzod qizi
student in Faculty of Foreign Language and Literature
at Renaissance University of Education
sevinch.yusupova@reu.edu
Karimova Nasiba Abdullo kizi
lecturer in English at Renaissance universityof education
nasibakarimova98@mail.com

Abstract: This article examines the rapid advancement of artificial intelligence and the risks associated with increasingly powerful systems. It explores AI "power" as a function of capability, including risks such as misinformation, system vulnerabilities, and autonomous decision-making. Based on a qualitative literature review, the study analyzes arguments for and against restricting access to advanced AI. The findings highlight controlled access as a balanced approach and emphasize the importance of governance, safety, and public accountability.

Keywords: artificial intelligence, AI risk, AI governance, controlled access, AI safety

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ НА ПОРОГЕ ВОЗМОЖНОСТЕЙ: РИСКИ, ДОСТУП И УПРАВЛЕНИЕ

Аннотация: В данной статье рассматривается стремительное развитие искусственного интеллекта и риски, связанные с ростом его возможностей. Анализируется понятие «силы» ИИ как функции его способности выполнять сложные задачи, включая распространение дезинформации, выявление уязвимостей систем и автономное принятие решений. На основе качественного анализа литературы исследуются аргументы за и против ограничения доступа к передовым системам ИИ. Результаты показывают, что контролируемый доступ является сбалансированным подходом и подчеркивают важность эффективного управления, безопасности и общественной ответственности.

Ключевые слова: искусственный интеллект, риски ИИ, управление ИИ, контролируемый доступ, безопасность ИИ

SUN'IY INTELLEKT QOBILIYAT CHEGARASIDA: XAVF, KIRISH IMKONIYATI VA BOSHQARUV

Annotatsiya: Ushbu maqola sun'iy intellektning tez sur'atlarda rivojlanishini va tobora kuchayib borayotgan tizimlar bilan bog'liq xavflarni tahlil qiladi. Unda sun'iy intellekt "kuchi" uning qobiliyati sifatida ko'rib chiqiladi, jumladan, dezinformatsiya tarqatish, tizim

zaifliklarini aniqlash va avtonom qaror qabul qilish kabi xavflar o‘rganiladi. Sifatli adabiyotlar tahliliga asoslangan holda, tadqiqot ilg‘or sun’iy intellektga kirishni cheklash va cheklamaslik bo‘yicha qarama-qarshi fikrlarni tahlil qiladi. Natijalar nazorat ostidagi kirish modelini muvozanatli yondashuv sifatida ko‘rsatadi hamda boshqaruv, xavfsizlik va jamoatchilik oldidagi javobgarlikning muhimligini ta’kidlaydi.

Kalit so‘zlar: sun’iy intellekt, AI xavfi, AI boshqaruvi, nazorat ostidagi kirish, AI xavfsizligi

In the past decade, artificial intelligence has progressed far faster than most experts anticipated. What once seemed like distant possibilities—machines producing natural conversations, assisting in drug discovery, or autonomously writing software—has now become reality. Today’s AI systems are no longer limited to narrow, specialized tasks; they are evolving into versatile tools that can perform a wide range of cognitive functions. This rapid transformation has sparked an important and somewhat unsettling question: could an AI become so advanced that making it publicly available might pose serious risks?

The concept of limiting access to powerful AI is not entirely new. From the early stages of modern machine learning development, researchers and organizations have sometimes chosen not to release their most capable models in full. The logic behind this caution is straightforward: technologies with great power often carry significant potential for misuse. Much like vulnerabilities in cybersecurity or sensitive discoveries in biochemical science, advanced AI systems—especially those capable of shaping information, automating complex decisions, or convincingly imitating human behavior—could be exploited on a large scale.

Literature Review

The concept of “power” in artificial intelligence (AI) is commonly framed not in terms of physical force, but in terms of capability and scope of influence. Recent scholarship highlights several domains in which advanced AI systems may pose significant risks when capabilities reach a certain threshold. These include the ability to generate highly convincing misinformation at scale, identify vulnerabilities in critical infrastructure, assist in the design of harmful biological or chemical agents, and influence human behavior through highly personalized persuasion. Additionally, concerns have been raised about systems that can operate autonomously in ways that are difficult to predict or control.

Existing studies suggest that many of these capabilities are already present in limited or fragmented forms. However, the primary concern emerges when these functionalities converge within a single, highly capable system that is widely accessible. This convergence amplifies both the scale and speed at which harm could occur, raising questions about the adequacy of current safeguards.

A substantial body of literature supports the argument for restricting access to advanced AI systems. Proponents emphasize that preventive measures are more effective than reactive ones, noting that once a system is publicly released, it cannot be withdrawn. Open access, they argue, may enable malicious actors to exploit AI technologies more rapidly than

governance frameworks can adapt. Furthermore, the issue of alignment—ensuring that AI systems operate in accordance with human values—remains an unresolved challenge. Systems that are insufficiently understood or lack robust control mechanisms may introduce unpredictable and potentially irreversible risks.

Conversely, critics of restriction highlight the risks associated with centralized control of powerful AI technologies. Concentrating access within a small number of organizations may lead to unequal distribution of benefits, reduced transparency, and potential misuse by corporate or governmental entities. Additionally, limiting access could hinder innovation by restricting collaboration and slowing the pace of scientific progress. Some scholars also question the effectiveness of restrictions, arguing that technological knowledge is inherently diffusive and that similar systems are likely to be developed independently across different regions.

In response to these competing perspectives, a growing consensus in the literature points toward a hybrid or "controlled access" model. This approach involves deploying advanced AI systems through monitored interfaces, where usage is regulated and safeguards are embedded. Such models aim to balance innovation and accessibility with oversight and risk mitigation, while also enabling researchers to study system behavior in real-world contexts before broader dissemination.Underlying this debate is a broader question of governance and trust. Key issues include determining who has the authority to define acceptable levels of risk, how access should be allocated, and how to ensure that decision-making processes serve the public interest rather than narrow institutional or competitive goals. These concerns are increasingly central as AI systems continue to evolve in capability and societal impact.

Methodology

This study adopts a qualitative, analytical approach to examine the concept of "power" in artificial intelligence and the associated debates حول access and restriction. The methodology is based on a structured review and synthesis of existing academic literature, policy reports, and industry practices related to advanced AI systems.First, relevant sources were identified through a targeted review of publications addressing AI risk, governance, alignment, and technological access. Second, a thematic analysis was conducted to categorize key arguments into two primary perspectives: the case for restricting access to advanced AI systems, and the case against such restrictions. Within each category, recurring themes—such as risk prevention, alignment challenges, centralization concerns, and innovation dynamics—were identified and analyzed.

Third, the study examined emerging industry practices, particularly the adoption of controlled access models. This involved analyzing how organizations implement monitoring mechanisms, usage limitations, and safety safeguards in real-world deployments of AI systems.

Finally, a comparative framework was used to evaluate the strengths and limitations of each perspective, with particular attention to the trade-offs between safety, accessibility, transparency, and innovation. By integrating insights from multiple sources and perspectives,

this methodology aims to provide a balanced and comprehensive understanding of the challenges associated with increasingly powerful AI systems and the implications of different access strategies.

Conclusion

The rapid development of artificial intelligence has transformed it from a narrow technological tool into a highly capable system with wide-ranging cognitive functions. This evolution has created both significant opportunities and serious challenges, particularly concerning the safe and responsible deployment of advanced AI systems. As the capabilities of AI continue to expand, concerns regarding misuse, unintended consequences, and lack of control become increasingly relevant.

This article has shown that the debate over whether advanced AI should be restricted is not straightforward. On one hand, unrestricted access may increase the risk of harmful applications, including misinformation generation, exploitation of system vulnerabilities, and autonomous actions that are difficult to predict or regulate. On the other hand, excessive restriction may concentrate power in the hands of a few organizations, limit transparency, slow down innovation, and create global inequalities in access to transformative technologies.

References:

1. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
2. Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258
3. Brundage, M., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. https://arxiv.org/abs/1802.07228
4. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
5. Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.