Секция «Технологии искусственного интеллекта в современной политике»

Культурная предвзятость в российских системах искусственного интеллекта: анализ идеологических репрезентаций в нейросети Шедеврум

Научный руководитель – Телин Кирилл Олегович

Саджая Лука Васильевич
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Факультет политологии, Кафедра государственной политики, Москва, Россия
E-mail: luka.sadjaya@yandex.ru

Проблема культурной предвзятости в алгоритмах становится ключевой в контексте разработки национальных систем Искусственного Интеллект (ИИ). Исследование фокусируется на нейросети Шедеврум (Яндекс) как примере, демонстрирующем противоречие между стремлением к технологическому суверенитету и доминированием западноцентричных культурных репрезентаций в российской разработке. Объект исследования — визуальные образы политических идеологий, генерируемые нейросетью, предмет — культурная предвзятость в их формировании. Методология включала систематический анализ изображений девяти идеологий (коммунизм, социализм, анархизм, национализм, фашизм, нацизм, либертарианство, либерализм, консерватизм) по четырём категориям: понятие, будущее, общество, типичный представитель.

Основные результаты исследования:

1. Нейросеть Шедеврум демонстрирует значительную западноцентричность в визуализации идеологий. Например, левые идеологии (коммунизм и социализм) представлены через стереотипные образы разрушения и упадка, с преобладанием красного цвета и военной символики. При этом коммунизм ассоциируется преимущественно с азиатскими и русскими персонажами, что отражает географическую привязку идеологии.

2. Анархизм визуализируется как постапокалиптический сценарий с разрухой и мародерами, что свидетельствует о смешении понятий "анархизм" и "анархия". Нацизм и фашизм представлены через депрессивные образы с элементами антиутопии, что также соответствует западным культурным стереотипам.

3. Либерализм и либертарианство ассоциируются с позитивными образами, связанными с американской государственной символикой, что подчеркивает доминирование западной перспективы в алгоритмах нейросети.

4. Консерватизм визуализируется через образы, напоминающие членов английского парламента или американских политиков, включая Дональда Трампа, что указывает на культурную предвзятость в репрезентации данной идеологии.

Проблема культурной предвзятости в нейросети Шедеврум отражает более широкий феномен "западного взгляда" в разработках искусственного интеллекта [4]. Этот феномен связан с концепцией "датаколониализма" [3]. Преобладание западных данных в обучающих выборках создает форму информационной колонизации, навязывая чуждые культурные модели. Аналогичные выводы представлены в исследованиях, изучающих навязывание стереотипов алгоритмами, в том числе в поисковых системамх и в других генеративных нейросетях [1, 2, 5, 6, 7, 8, 9]. Эти исследования подтверждают, что алгоритмы искусственного интеллекта часто становятся инструментами культурной гегемонии, что проявляется и в работе нейросети Шедеврум.

Выводы:

1. Шедеврум воспроизводит западноцентричные идеологические стереотипы, что противоречит целям технологического суверенитета.

2. Для создания культурно-релевантных систем ИИ необходимы датасеты, отражающие российские интерпретации социально-политических концепций.

3. Разработка таких систем должна избегать идеологической ангажированности, стремясь к отражению многообразия общественных дискурсов.

Исследование подчёркивает актуальность коррекции культурной предвзятости в отечественных нейросетях как условия достижения подлинного технологического суверенитета.

Источники и литература

1) Basu, A., Babu, R. V., & Pruthi, D (2023). Inspecting the Geographical Representativeness of Images from Text-to-Image Models. Indian Institute of Science, Bangalore.
2) Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, 1–15.
3) Couldry, N., Mejias, U. A. (2019). The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford University Press.
4) Gautam, S., Venkit, P. N., Ghosh, S. (2024). From Melting Pots to Misrepresentations: Exploring Harms in Generative AI.
5) Hagerty, A., Rubinov, I. (2019). Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence.
6) Kidd, C., Birhane, A. (2023). How AI Can Distort Human Beliefs. Science, 380(6651), 1222–1223
7) Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press, 2018.
8) Biases in large image-text AI model favor wealthier, Western perspectives // University of Michigan News. 2023. URL: https://news.engin.umich.edu/2023/12/biases-in-large-image-text-ai-model-favor-wealthier-western-perspectives/ (дата обращения: 05.03.2025).
9) Rest of World. 2023. URL: https://restofworld.org/2023/ai-image-stereotypes/ (дата обращения: 05.03.2025).