Стребкова Карина

Artificial Central Intelligence Agency?
Риски и перспективы использованиягенеративного ИИ в разведывательном сообществе

Тестирование генеративного ИИ в ЦРУ

В последнее время наблюдается значительное развитие и использование генеративного искусственного интеллекта (ИИ) в различных сферах человеческой деятельности. Одной из областей, где его применение может иметь особое значение, является разведывательные службы. Например, недавно выяснилось, что ЦРУ (Центральное Разведывательное Управление, США) планирует исследовать потенциал генеративного ИИ для поддержки своей миссии. «Честно говоря, мы видели ажиотаж в социуме вокруг ChatGPT. Это, безусловно, переломный момент в этой технологии, и нам определенно нужно [изучить] способы, которые позволят нам использовать новые технологии. И я думаю, что нам нужно выработать дисциплинированный подход к этому»[1] [1], – заявила Лакшми Раман, директор подразделения ЦРУ поИИ, раскрывая планы агентства на ежегодном саммите по искусственному интеллекту Потомакского клуба офицеров [2]. Разведка США и ранее искала возможности для применения разработок в области ИИ: ещё со времён холодной войны [3], однако тогда не было такого «переломного момента в этой технологии». Какие риски и перспективы здесь скрываются?

Аналитика и превентивные меры

«При правильном использовании искусственный интеллект может находить закономерности и тенденции в огромных объемах открытых источников и тайно добытых данных, что не под силу человеческому разуму... <…> Наши противники быстро используют информацию из открытых источников, и мы должны делать это быстрее и лучше, чем они» [4], – сказал директор ЦРУ Уильямс Бёрнс на ежегодной лекции в Дитчли, организованной внешнеполитической группой Ditchley Foundation (Великобритания). Действительно, ИИ как инструмент, в том числе большие языковые модели, может быть полезен для анализа больших массивов данных, которые невозможно обработать одному человеку, даже если он потратит на это всю жизнь. Как прогнозирует IDC, к 2025 году только один Интернет вещей (IoT) будет генерировать почти 80 миллиардов зеттабайт данных [5]. Соответственно, с ростом цифрового пространства увеличивается не только объём данных, но и количество информации, в связи с чем возникает проблема грамотной реализации цикла Data Science (см. рис.), где будет сбалансированное сочетание технологий и возможностей человеческого интеллекта, чтобы принять адекватное решение актуальных задач.

---

[1] *«Honestly, we've seen the excitement in the public space around ChatGPT. It's certainly an inflection point in this technology, and we definitely need to [be exploring] ways in which we can leverage new and upcoming technologies. And I think the way we're approaching it is we need to approach it in a disciplined way»*

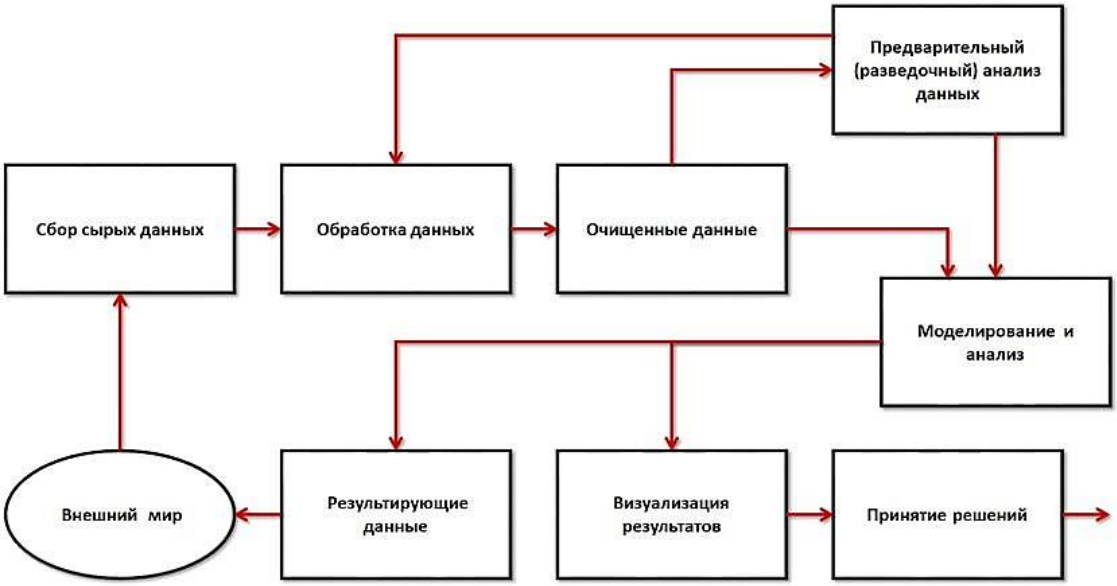

Рисунок – Технологический цикл Data Science [10]

Как указывает Уильям Бёрнс, одной из приоритетных стратегических задач ЦРУ является совершенствование технологического цикла анализа данных и информации, чтобы выявлять скрытые закономерности, что может помочь в раннем обнаружении и предсказании потенциальных проблем, а также реакции на неё [4]. Кроме того, это позволит аналитикам разведки больше «сосредоточиться на том, что они умеют делать лучше всего: давать обоснованные суждения и выводы о том, что наиболее важно для политиков и что наиболее значимо для наших интересов»[2] [4].

Таким образом, учитывая позиции руководства ЦРУ, развивающиеся технологии ИИ могут быть чрезвычайно полезны для аналитики. Однако некоторые инструменты ИИ, в частности большие языковые модели, могут нанести вред именно этой области разведки, поскольку генеративный ИИ имеет возможность самостоятельно генерировать контент на основе обучения из большого объема данных. То есть языковая модель может быть применена для успешной имитации «человеческого» текста, но лежащие в его основе факты могут оказаться фальшивыми [3]. Это явление среди специалистов получило название «галлюцинации». Склонность неспециально дезинформировать пользователей создаст серьёзные проблемы в сфере разведки: одно дело, когда ChatGPT представляет несуществующие рестораны в Москве, и совсем другое, если обученная на секретной информации модель «выдумывает» такие закономерности, которые приводят аналитиков к неправильному выводу о состоянии ядерной программы Ирана. Ситуация усложняется тем, что модели не дают сведения об источниках информации или о том, как она была создана. Однако всё это не говорит о том, что применение языковых моделей бесполезно для разведки.

Это побуждает к тому, как выразилась Лакшми Раман, к разработке «дисциплинированного подхода» к использованию генеративного ИИ. Разведывательному сообществу ещё предстоит создать новые процедуры анализа информации и контроля ИИ,а также технологии, которые позволят минимизировать галлюцинации нейросети.

---

[2] *«…freeing up our officers to focus on what they do best: providing reasoned judgments and insights on whatmatters most to policymakers, and what means most for our interests»*

Предвзятость генеративного ИИ

Ещё одним неочевидным вызовом для разведывательного сообщества является предвзятость языковых моделей. Информация, включенная в оцифрованные массивы для генеративного ИИ, создавалась людьми, которые склонны искажать данные и иметь разного рода предубеждения. Неудивительно, что нечто подобное наблюдается при принятии решения нейросетью. Л. Лян и Д.Е. Акуна в своём исследовании восприятия предубеждений в моделях ИИ [6] указывают, что предубеждения находятся очень глубоко в моделях глубокого обучения и создают решения, которые дискриминируют людей по полу и другим социальным статусам, а также не совсем корректно могут делать предсказания. Иными словами, ИИ становится чёрным ящиком, где пути анализа информации совершенно неочевидны. Схожим образом работает человеческое сознание при обработке информации. Это значит, как говорят Л. Лян и Д.Е. Акуна, что можно адаптировать методы экспериментальной психологии для выявления предубеждений в моделях ИИ, после чего использовать когнитивно-поведенческую терапию для коррекции процессов внутри ИИ.

Сталкивалось ли разведывательное сообщество с подобным ранее? Да, при исследовании ограничений когнитивной системы человека при анализе разведданных. Одним из пионеров в этой области является Ричард Хойер, аналитик ЦРУ, известный своей работой «Psychology of Intelligence Analysis» [7]. В ней подробно описано, каким образом сложившиеся ментальные модели, когнитивные искажения и предубеждения, а также работа когнитивных механизмов в целом мешают грамотному анализу и принятию решений. Как структурировать аналитический процесс, чтобы избежать хотя бы части ограничений, рассказано в работе «Structured Analytic Techniques for Intelligence Analysis» [8]. Таким образом, разведывательное сообщество активно использует теоретические концепции и прикладные исследования в области когнитивистики для совершенствования аналитики, которая создаётся людьми. Возможно, что к середине XXI века подобные работы потребуются при обучении и использовании генеративных моделей, в том числе в разведке.

Ответственность и власть

Помимо технических сложностей использования генеративного ИИ в разведывательном сообществе, существуют проблемы иного характера.

Как уже было указано ранее, генеративный ИИ имеет свойство создавать, модифицировать и подделывать информацию. Некорректное и безответственное использование генеративного ИИ на его текущем этапе развития может подорвать доверие к правительству среди граждан и формированию плохой репутации разведывательных служб.

Кроме того, с появлением в разведывательном сообществе такого инструмента как генеративный ИИ, открываются невероятные возможности для шпионажа и злоупотребления властью, так как созданные алгоритмы могут быть использованы для манипулирования информацией, подавления политических оппонентов и кражи секретных ведений. Так, по данным документа «Artificial Intelligence and National Security» Гарвардской школы Кеннеди [9], китайское правительство взломало многие американские оборонные и военные структуры, связанные с реализацией секретной

программы F-35, и приобрело почти всю интеллектуальную собственность, которая касалась разработки этого самолёта. Кроме того, есть предположение, что Китай похитил при помощи кибершпионажа чрезвычайно важную информацию, связанную с ядерным арсеналом США. Не уточняется, какой именно тип ИИ был задействован при этих операциях, но ясно, что отдельные инструменты генеративного ИИ могут быть использованы для более оперативного считывания сведений из большого объема текстовых данных.

Заключение

Внедрение генеративного ИИ в разведывательное сообщество является двуединым процессом с различными рисками и перспективами. Хотя использование генеративного ИИ может способствовать решению сложных задач и повышению безопасности общества, существуют риски злоупотребления властью, недостатка ответственности и угрозы кибербезопасности. Поэтому для обеспечения эффективного и этичного использования генеративного ИИ в разведывательном сообществе, необходимо разработать специальные, возможно, международные нормы, гарантирующие защиту прав и свобод граждан, а также новые технологии и способы анализа информации, способные минимизироватьриски применения генеративного ИИ.

Список использованных источников

1. Vincent B. CIA to investigate how generative AI (like ChatGPT) can assist intelligence agencies / DefenseScoop. – February 16, 2023. – URL: https://defensescoop.com/2023/02/16/cia-to-investigate-how-generative-ai-like-chatgpt-can-assist-intelligence-agencies/ (дата обращения: 16.08.2023).
2. Potomac Officers Club. Lakshmi Raman, Director of Artificial Intelligence Central Intelligence Agency (CIA) / Potomac Officers Club. – URL: https://potomacofficersclub.com/speakers/lakshmi-raman/ (дата обращения: 16.08.2023).
3. Townley D. Intelligence agencies have used AI since the cold war – but now face new security challenges / The Conversation. – May 3, 2023. – URL: https://theconversation.com/intelligence-agencies-have-used-ai-since-the-cold-war-but-now-face-new-security-challenges-204320 (дата обращения: 16.08.2023).
4. Burns, J. W. A World Transformed and the Role of Intelligence / 59th Ditchley Annual Lecture. – July 1, 2023. – URL: https://www.ditchley.com/sites/default/files/Ditchley%20Annual%20Lecture%202023%20transcript.pdf (дата обращения: 16.08.2023).
5. IDC. Future of Industry Ecosystems: Shared Data and Insights / IDC Blog. – January 6, 2021. – URL: https://blogs.idc.com/2021/01/06/future-of-industry-ecosystems-shared-data-and-insights/ (дата обращения: 16.08.2023).
6. Liang L. & Acuna D. E. Artificial mental phenomena: Psychophysics as a framework to detect perception biases in AI models / Proceedings of the 2020 conference on fairness, accountability, and transparency. – 2020. – pp. 403-412.
7. Heuer R. J. Psychology of intelligence analysis / Washington, DC: U.S. Central Intelligence Agency, Center for the Study of Intelligence, 1999.
8. Pherson, R. H. & Heuer R. J. Structured analytic techniques for intelligence analysis. / Cq Press, 2020.
9. Allen G. & Chan T. Artificial Intelligence and National Security / Belfer Center for Science and International Affairs, Harvard Kennedy School. – 2017. – URL: https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20%20final.pdf (дата обращения: 16.08.2023).
10. Черняк Л. Данные / TAdviser. – 6 июня 2022. – URL: https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%94%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5 (дата обращения: 16.08.2023).