Секция «21.2 Международная безопасность: новые и традиционные вызовы и угрозы»

Анализ конфликта департамента обороны США и Anthropic: сложности в использовании технологий генеративного ИИ в целях национальной безопасности

Научный руководитель – Малов Андрей Юрьевич

Шмаков Георгий Андреевич

Студент (магистр)

Московский государственный университет имени М.В.Ломоносова, Факультет мировой политики, Кафедра международной безопасности, Москва, Россия

E-mail: shmakovga@yandex.ru

27го февраля этого года, президент США Дональд Трамп дал команду федеральным органам власти прекратить использование ИИ-услуг компании Anthropic [1]. В этот же день, секретарь обороны США Питер Хегсет назначил Anthropic "риском цепочки поставок", запрещая подрядчикам ОПК использовать продукты и услуги этой компании в работах, связанных с департаментом обороны [2]. 28го февраля этого года, Anthropic заявила о намерении оспорить это решение в суде.

Это решение происходит из растущего внимания к использованию больших языковых моделей (LLM) в вооруженных конфликтах. ИИ-продукты Anthropic (модели Claude) использовались в операции Абсолютная Решимость в Венесуэле [3], а также в операции Эпическая Ярость в Иране [4]. Использование генеративных ИИ-технологий на поле боя теперь является реальностью, и одним из непредвиденных вызовов этой реальности оказывается конфликт между оперативными требованиями вооруженных сил, национальными интересами государства и правительства, правилами ИИ-компаний в области этики, и технической реальностью искусственного интеллекта.

Согласно анонимным источникам, требования Пентагона к системам Anthropic сводятся к следующим:

1. "Неограниченное военное использование" больших языковых моделей Anthropic [5];
2. Идеологическая предвзятость моделей Claude [2].

Конфликт возникает в связи с этическими правилами Anthropic, связанными с использованием их моделей для автономного вооружения и системами массового наблюдения на территории США. [5] Важно отметить, что это ограничение не является новым, и происходит из моральных соображений ИИ-разработчиков – так, почти 10 лет назад, в связи с внутренними протестами, корпорация Alphabet вышла из проекта Мавен, раннего проекта по использованию ИИ в ВС США. Протесты были вызваны схожими соображениями. [6]

Также, интересно отметить, что Пентагон готов к переговорам с ИИ-компаниями – так, генеральный директор OpenAI Сэм Олтман, в твите от 28го февраля этого года, упомянул эти же соображения, но заявил о готовности к сотрудничеству с Пентагоном. [7]

Рассмотрим более пристально требования Пентагона к моделям Anthropic – Пентагон выдвинул два требования, но они взаимосвязаны, и связаны также с другими трендами в области информационных технологий в обороне.

Первое требование – неограниченное военное использование моделей. Это требование можно рассматривать как параллельное требованиям вооруженных сил различных стран мира к полному контролю над их вооружениями. В области информационных технологий, с учетом технологий телеметрии и контроля доступа, это требование иногда оказывается под угрозой. Так, 15го февраля этого года, государственный секретарь по обороне Нидерландов Гийс Туинман заявил в интервью, что истребители 5го поколения F-35 могут быть "взломаны" для получения неограниченного доступа к ним [8]. Это заявление показывает схожую озабоченность ВС Нидерландов технологической зависимостью от США, и таким образом, способностью США ограничить независимые действия ВС Нидерландов.

Второе требование, требование по идеологической непредвзятости моделей Claude, не имеет параллелей в сфере информационных технологий, и связано с тем, что большие языковые модели имеют вероятностный характер, и слишком сложны для их полноценного понимания. Согласно меморандумам Пентагона, департамент обороны США опасается "идеологического тюнинга, предотвращщающего их [больших языковых моделей] способность предоставлять объективно правдивые ответы на запрос пользователя". [9] Если довести аргумент до его логического конца, департамент обороны США боится, что идеологическая предвзятость, заложенная в модель, может скрытным образом повлиять на ход боевых действий, противореча интересам командования.

Важно отметить, что оба эти требования к ИИ-технологиям важны в даже большей степени для стран, не имеющих собственных ИИ-флагманов – то есть всех стран, кроме США, и, в меньшей степени, Китая.

Для обоих из требований Пентагона, существуют технические методы их выполнения:

• Касаемо требования о неограниченном использовании моделей, существует система размещения ПО под контролем заказчика ("on premises") – так, Google предоставляют подобный доступ к своим флагманским моделям. [10]
• Касаемо идеологической предвзятости, существует технология "дообучения" (fine-tuning) больших языковых моделей, способная изменить или убрать их предвзятость. [11]

Ключевой сложностью в данном случае является вопрос набора данных для дообучения, и метрик для оценки дообученных моделей – для получения контроля над поведением модели, каждому государственному актору потребуется свой набор данных и свои метрики, соответствующие их задачам и их видению непредвзятости.

Источники и литература

1) Guardian: Трамп приказывает департаментам США прекратить использование технологий Anthropic в виду диспута об этике ИИ: https://www.theguardian.com/us-news/2026/feb/27/trump-anthropic-ai-federal-agencies
2) The Hacker News: Пентагон объявляет Annthropic риском для цепочки поставок в связи с диспутом ИИ и вооруженных сил: https://thehackernews.com/2026/02/pentagon-designates-anthropic-supply.html
3) Guardian: ВС США использовали ИИ-модель Anthropic, Claude, в рейде на Венесуэлу, заявляет репортаж: https://www.theguardian.com/technology/2026/feb/14/us-military-anthropic-ai-model-claude-venezuela-raid

4) StarAdvertiser: США используют ИИ Anthropic, бомбардировщики B-2, дроны-камикадзе, в ударах по Ирану: https://www.staradvertiser.com/2026/03/02/breaking-news/u-s-uses-anthropic-ai-b-2-bombers-suicide-drones-in-iran-strikes/

5) Associated Press: Генеральный директор Antropic говорит, что компания “не может с честной совестью принять” требования Пентагона по использованию ИИ: https://apnews.com/article/anthropic-ai-pentagon-hegseth-dario-amodei-9b28dda41bdb52b6a378fa9fc80b8fda

6) Bloomberg: Google отрекается от ИИ-вооружений, но все еще продолжит сотрудничать с вооруженными силами: https://www.bloomberg.com/news/articles/2018-06-07/google-renounces-ai-for-weapons-but-will-still-sell-to-military

7) Официальный твит Сэма Олтмана: https://x.com/sama/status/2027578580159631610

8) Clash Report: Голландский чиновник предлагает “взлом” F-35: https://clashreport.com/defense/articles/dutch-official-floats-f-35-jailbreak-n1xyjxiieo

9) Меморандум департамента обороны США: Стратегия искусственного интеллекта для Департамента Войны: https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF

10) Google Cloud: приносим Gemini и Google Agentspace к вам on-premises: https://cloud.google.com/blog/products/ai-machine-learning/run-gemini-and-ai-on-prem-with-google-distributed-cloud

11) У Сяо-Кун, Чэнь М., Ли Ваньи, Ван Жуй, Лиу Цзя, Хван К., Хао Исюэ, Пан Яньжу, Мэн Цингуо, Хуан Кайбин, Ху Лонг, Гуизани М., Чао Наипен, Фортино Дж., Лин Фэй, Тянь Ёнлинь, Ниято Дусит, Ван Фэй-Юэ. Понятие, возможности и проблемы тонкой настройки больших языковых моделей // Big Data and Cognitive Computing. 2025. Т. 9, № 4. Статья 87. DOI: 10.3390/bdcc9040087.