

отношении взаимодействия девиантов с окружающей социальной средой.

Таким образом, важным направлением социальной политики белорусского государства на ближайшие десятилетия является активизация и совершенствование социально-правовой системы предупреждения преступности, пьянства, наркомании и других антиобщественных явлений. Эффективная деятельность по предупреждению антиобщественных явлений обуславливает актуальную социальную потребность в развитии криминологического и другого антидевиантного законодательства. Необходимы следующие новые Законы Республики Беларусь: «Об основах государственной антидевиантной политики»; «Предупреждение преступности, пьянства, наркомании

и других антиобщественных явлений»; «Предупреждение правонарушений несовершеннолетних и основы ювенального правосудия»; «Виктимологическая политика, защита прав и интересов жертв преступлений и других правонарушений». Практическая реализация предлагаемых теоретических, законопроектных и правоприменительных инноваций будет способствовать формированию в стране современной социально-правовой системы предупреждения антиобщественных явлений, обеспечит устойчивое снижение распространенности и негативных последствий антиобщественных явлений и тем самым содействовать социально-экономическому, демографическому, социальному и культурному развитию современного белорусского общества.

Ильиных Н.А.

Сибирский юридический институт МВД России (г. Красноярск)

Угрозы применения больших языковых моделей в преступной деятельности

Сегодня все чаще в различных областях науки встречается термин «большие языковые модели»¹. Наиболее корректно говорить, что он охватывается более широким понятием – «большие генеративные модели», под которым согласно Национальной стратегии развития искусственного интеллекта на период до 2030 года понимаются: «модели искусственного интеллекта, способные интерпретировать (предоставлять информацию на основании запросов, например, об объектах на изображении или о проанализированном тексте) и создавать мультимодальные данные

(тексты, изображения, видеоматериалы и тому подобное) на уровне, сопоставимом с результатами интеллектуальной деятельности человека или превосходящем их»². В свою очередь, большие языковые модели обрабатывают исключительно текстовые данные.

Такие популярные модели как YandexGPT, GigaChat, ChatGPT, DeepSeek и др. плотно вошли в нашу повседневную жизнь, помогая решать рабочие и бытовые задачи. По состоянию на конец 2025 года согласно данным социологических опросов, свыше половины россиян как минимум пробовали ис-

1 Зайцев Д.В. Почему большие языковые модели не (всегда) рассуждают как люди? // Вестник Московского университета. 2024. № 1. С. 78.

2 О развитии искусственного интеллекта в Российской Федерации : Указ Президента РФ от 10.10.2019 № 490.

пользовать одну или несколько из них¹. Вместе с тем с ростом популярности среды населения, возникают и новые угрозы, связанные с их эксплуатацией в криминальных целях.

В качестве наиболее актуальных направлений использования больших языковых моделей для совершения преступлений можно выделить: генерацию качественных фишинговых писем на разных языках, обладающих индивидуальными особенностями, что позволяет им вызывать доверие у пользователей и обходить фильтры, настроенные на выявление спама²; разработка и эксплуатация вредоносного программного обеспечения; помощь в создании оружия³; генерация текстов экстремистского содержания и др.

Возможности применения больших языковых моделей в криминальных целях ограничиваются лишь инструкциями о соответствующих запретах, заложенными в них разработчиками в ходе обучения. К счастью, если просто попросить наиболее популярные большие языковые модели написать инструкцию по изготовлению взрывчатки, то они ответят отказом. Однако данная технология весьма восприимчива к промптам (текстовым запросам пользователя, от англ. *prompt* – «побуждать», «подстрекать», «подсказывать»), при помощи которых можно обойти защитные инструкции, заложенные разработчиками и добиться генерации противоправного контента. Данная проблема получила самостоятельное обозначение – «*jailbreaking*» (англ. – «побег из тюрьмы»)⁴.

Сегодня создатели большинства языковых моделей, принадлежащих государственному или крупным коммерческим организациям, постоянно совершенствуют защитные инструкции, противодействуя вышеуказанной угрозе, однако сложно говорить о возможности ее полного устранения в будущем, поскольку разработчики часто находятся в позиции «догоняющих», по отношению к преступникам. Вместе с тем здесь мы говорим о случаях, когда исходный код данных моделей часто недоступен для модификации пользователям и принадлежит конкретным физическим или юридическим лицам, которые действуют в рамках правового поля и активно сотрудничают с правоохранительными органами, стараясь минимизировать эксплуатацию их продуктов в преступных целях. Гораздо менее однозначной в данной связи видится ситуация, обусловленная распространением больших языковых моделей с открытым исходным кодом, который может использовать абсолютно любой пользователь.

В силу сложности разработки до недавних пор доступ к рассматриваемым технологиям доступ был крайне ограниченным. Однако в последние годы ситуация стала меняться в связи с ростом доступности вычислительных мощностей, совершенствованием подходов к обучению генеративных моделей и появлением в открытом доступе больших и качественных датасетов. Сейчас только в одном из агрегаторов любому пользователю доступно для неограниченного использования более 322 тысяч готовых моделей, способных обрабатывать текст

1 Нейросети: инструмент, а не магия // ВЦИОМ. URL: <https://wciom.ru/analytical-reviews/analiticheskii-obzor/neiroseti-instrument-a-ne-magija> (дата обращения: 25.01.2026).

2 Донских Н.И. Искусственный интеллект в киберугрозах: возможности и вызовы использования нейронных сетей для фишинговых атак и их детектирования // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические Науки. 2025. № 01. С. 71.

3 Пинкевич Т.В., Конев Д.А. Риски использования искусственного интеллекта как орудия и средства совершения преступлений: проблемы правового регулирования и противодействия // Правопорядок: история, теория, практика. 2024. № 4. С. 151.

4 Крохин А.С., Гусев М.М. Анализ влияния обфускации входных данных на эффективность языковых моделей в обнаружении инъекции подсказок // Программные системы и вычислительные методы. 2025. № 2. С 44-62.

и свыше 414 тысяч готовых обучающих выборок¹. Запуск таких моделей все еще часто требует от пользователя наличия специальных навыков и значительных вычислительных мощностей, однако квалификационный порог постоянно снижается, а соответствующие технические средства можно приобрести в весьма доступном бюджете или воспользоваться многочисленными облачными решениями.

Безусловно, широкое распространение моделей с открытым исходным кодом влечет за собой массу положительных эффектов, помогая, например, преодолевать одну из таких современных проблем технологий искусственного интеллекта, как его непрозрачность². Однако нельзя не обращать внимание и на возникающие в данной связи криминальные угрозы.

Получая доступ к исходному коду, пользователь, обладающий необходимыми техническими навыками, может самостоятельно изменять любую модель, обучая ее генерировать в том числе и противоправный контент. При этом, как отмечают в своем исследовании А.А. Пичугов, Д.Е. Намиот и Е.В. Зубарева, сейчас процесс обучения значительно упростился, за последние 4 года в данной области произошел «технологический сдвиг от подходов обучения, требующих десятков и сотен тысяч примеров, к методикам, демонстрирующим впечатляющую эффективность при использовании минимального количества данных, – вплоть до одного-единственного примера»³.

В результате появилось множество открытых больших языковых моделей, у которых отсутствуют какие-либо защитные инструкции и внутренние ограничения на генерацию контента. Они часто имеют в своем названии пометку «uncensored» (англ. – «без цензуры») и пользуются большой популярностью среди пользователей. Данные модели нередко могут ответить на вопросы о том, как изготовить наркотическое средство или как правильно уничтожать следы после совершенного убийства. Несмотря на порой низкое качество ответов, потенциал для их улучшения сегодня практически ничем не ограничен. В данной сфере уже существуют и крупные коммерческие проекты, которые прямо продвигают продукты, генерирующие, например, вредоносный код⁴. Сейчас не составляет особого труда найти их в сети в совершенно открытом доступе.

Таким образом, широкое распространение больших языковых моделей, помимо очевидно положительных перспектив, влечет за собой и криминальные угрозы. Считаем, что проблема требует дальнейшей научной разработки и правовой регламентации. На наш взгляд, наиболее перспективным направлением противодействия данной угрозе видится в совершенствовании механизмов контроля со стороны правоохранительных органов над созданием коммерческими организациями новых больших языковых моделей с закрытым исходным кодом, а также выявления и ограничения доступа к открытым моделям, обученным для использования в преступных целях.

1 Hugging Face: The AI community building the future // Hugging Face. URL: <https://huggingface.co> (дата обращения: 04.02.2026).

2 Дремлюга Р.И. Использование искусственного интеллекта в преступных целях: уголовно-правовая характеристика // Азиатско-тихоокеанский регион: экономика, политика, право. 2021. № 3. С. 153-165.

3 Пичугов А.А., Намиот Д.Е., Зубарева Е.В. Современные методы обучения больших языковых моделей с минимумом данных: От одного примера к абсолютному нулю – академический обзор // International Journal of Open Information Technologies. 2025. № 6. С. 114-123.

4 «Злые двойники ChatGPT»: эксперты об угрозах и развитии искусственного интеллекта // Татар-информ. URL: <https://www.tatar-inform.ru/news/zye-dvoiniki-chatgpt-eksperty-ob-ugrozax-i-razvitii-iskusstvennogo-intellekta-5934511> (дата обращения: 15.01.2026).