ПРОБЛЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КОНТЕКСТЕ РЕЖИМОВ ЗАПРЕТА ХИМИЧЕСКОГО И БИОЛОГИЧЕСКОГО ОРУЖИЯ

Искусственный интеллект (ИИ) способен коренным образом изменить характер научной, экономической и социальной деятельности в различных отраслях. В то же время вероятность негативных последствий использования ИИ по двойному назначению порождает все больше опасений. Угрозы являются как реальными, так и потенциальными и проистекают из существующих и потенциальных возможностей искусственного интеллекта.

Озабоченность по поводу двойного применения ИИ также была выражена в контексте режимов запрещения химического и биологического оружия (ХБО). В частности, тревогу вызывают возможности ИИ в области проектирования, разработки, размещения и обнаружения ХБО.

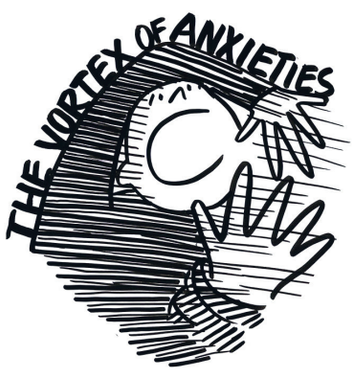
Общественность все больше волнуют предположения о том, что ИИ может способствовать разработке сверхтоксичных ОВ или предоставить государственным и негосударственным субъектам недорогие планы разработки и применения ХБО.

В настоящей аналитической статье рассмотрены четыре потенциальные проблемы ИИ и ХБО, а также предложены пути решения этих проблем на основе существующих этических принципов ответственного применения ИИ.

Определив, с какими негативными свойствами ИИ призваны бороться данные принципы, можно лучше понять, как эти свойства ИИ могут порождать конкретные проблемы в рамках режимов запрещения ХБО. Такой упреждающий подход может использоваться в рамках текущей работы по выявлению природы потенциальных проблем.

Ключевые тезисы

- ИИ существует не сам по себе, но является частью более крупной системы и применяется совместно с другими технологиями для улучшения обработки данных и принятия решений.
- ИИ может приводить к возникновению новых проблем и опасений, однако многие из них принципиально не отличаются от тех, что связаны с другими технологиями или практиками. ИИ может усугубить и изменить характер существующих проблем, и поэтому его необходимо рассматривать в более широком контексте с учетом систем, в которых он работает.
- Для смягчения потенциально опасных последствий применения ИИ не требуется полной смены архитектуры управления. Выявление и решение потенциальных проблем, связанных с ИИ, может помочь в изменении и дополнении существующих структур.



UNIVERSITY OF SUSSEX

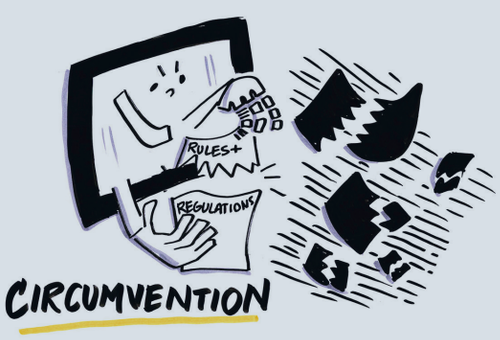
HSP | THE HARVARD SUSSEX PROGRAM

ИИ + (ТЕХНОЛОГИЯ)

Интеграция ИИ с другими технологиями создает серьезные проблемы для режимов запрещения ХБО. Тревожные сценарии учитывают возможность ИИ увеличивать масштаб и степень рисков, порождаемых существующими технологиями двойного назначения. Например, использование искусственного интеллекта в автономных транспортных средствах для доставки химических средств борьбы с беспорядками (ХСББ) может вызвать дополнительные сомнения в безопасности и законности применения ХСББ в соответствии с КЗХО. Кроме того, использование ИИ в сочетании с робототехникой нового поколения позволяет разрабатывать и синтезировать химические или биологические вещества дистанционно и без привлечения значительных человеческих ресурсов. Этические принципы в этом случае могут в большей степени основываться на ценностях «Безопасность и надежность» и «Ориентированность на человека», а инфраструктура, управление и оценка намерений могут помочь прояснить потенциальные меры управления использованием ИИ.

ПУТИ ОБХОДА ОГРАНИЧЕНИЙ

ИИ может помочь злоумышленникам преодолеть традиционные ограничения, препятствия и меры контроля, мешающие разработке, хранению и размещению ХБО. Например, предложить способы обхода контроля импорта/экспорта веществ путем выявления или моделирования не включенных в перечень прекурсоров для химического синтеза. Кроме того, ИИ позволяет сократить количество человеческих ресурсов, необходимых для разработки или размещения оружия, и тем самым снизить затраты и избежать потенциальных «моральных дилемм». Наконец, ИИ может быть использован для разработки веществ, трудно поддающихся обнаружению, от которых может быть сложнее защититься. В свете этого принципы «Контролируемость» и «Подотчетность» диктуют необходимость изучения роли управляющих структур в разработке ИИ.



ИИ ДОЛЖ

Прозрачным и подотчетным

«Возможность объяснить, как и почему были достигнуты конкретные результаты. Исходные данные, структура проекта и операционные системы, процесс работы в целом и результаты должны быть прозрачными и подлежать отчетности».

Контролируемым

«Возможность контролировать исходные данные, вычисления и результаты, чтобы перенаправлять, изменять, отменять или останавливать операции; результаты также должны быть потенциально контролируемыми (для предотвращения автоматического распространения)».

ЕН БЫТЬ...

Ориентированным на человека

«Ответственность за полный цикл разработки и эксплуатации ИИ и использование его результатов несут люди. Понимание того, где и как они задействованы в этом процессе, может помочь в оценке их намерений».

Безопасным и надежным

«Для защиты ИИ от действий злоумышленников необходимо использовать системы кибер- и инфраструктурной безопасности с возможностью перехода в аварийный режим, остановки/отмены действий и запуска механизмов проверки, позволяющих убедиться, что результаты работы ИИ не представляют опасности для людей».

ПОСТЕПЕННАЯ ЛЕГИТИМИЗАЦИЯ

Намеренное нарушение правовых запретов и социальных табу в рамках проведения определенных исследований и разработок — далеко не новая проблема. Однако использование потенциала ИИ для упрощения и ускорения научно-исследовательских процессов и создания новых направлений работы, легко нарушающих границы законности, может предоставить злоумышленникам новые возможности для разработки ОВ и технологий, способных поставить под удар договоренности в отношении ХБО. Например, использование ИИ в свете существующих допущений может привести к непредусмотренным или сомнительным видам применения ХСББ, выходящим за пределы правового поля. В то же время военные исследования биологических процессов, ставящие своей целью возможность ведения войны без человеческих потерь, могут выйти на новый уровень благодаря технологиям ИИ. В данном случае принципы «Подотчетность» и «Ориентированность на человека» могут дать представление об управлении различными субъектами (или управляющей роли субъектов) для сохранения правовых и социальных границ с учетом новых возможностей.

ЧЕРНЫЙ ЯЩИК

Непрозрачный характер исходных данных, методов и результатов работы ИИ налагает дополнительные обязательства на разработчиков и пользователей. Неосязаемость, непрозрачность и малопонятность систем ИИ будет являться источником проблем во многих областях. Например, при доказательстве фактов неправомерного использования ИИ может быть трудно отследить необходимые данные, если применялся готовый инструмент ИИ, о точном происхождении, обучающих данных и методах дедукции которого мало что известно. Кроме того, выявление субъектов, использующих инструменты ИИ для разработки и производства потенциальных видов ХБО может быть затруднено в ситуациях, когда доступ к необходимым для этого инструментам и данным имеет множество лиц. В этих случаях на первый план выходят принципы «Прозрачность», «Подотчетность» и «Контролируемость», предполагающие повышение прозрачности и обмен информацией для контроля за созданием и использованием алгоритмов ИИ.

Выводы

Описанные четыре проблемы ИИ демонстрируют, что внедрение технологий ИИ не столько порождает новые категории угроз и рисков, сколько дополняет существующие. Важно отметить, что конкретные сценарии могут подходить под несколько категорий угроз в зависимости от того, как ИИ взаимодействует с существующими технологиями и способами управления. Таким образом, проблемы ИИ следует рассматривать в рамках контекста и проблем, уже знакомых субъектам, действующим в режимах запрещения ХБО. В этих контекстах влияние ИИ, вероятно, будет заключаться в следующем:

- Ускорение процессов исследований и разработок;
- Открытие новых перспективных направлений исследований;
- Уменьшение прозрачности;
- Усложнение способов предоставления информации, ее значения и последствий использования.

На следующем этапе необходимо оценить, способны ли существующие формы управления, благодаря имеющимся возможностям или при условии внесения изменений, уменьшить влияние ИИ на существующие и возникающие проблемы.

Четыре этических принципа ответственного ИИ указывают на то, какие вопросы мы должны задавать в рамках работы по выявлению, описанию и минимизации негативных последствий применения ИИ.

Дополнительная информация:

Данное исследование проводилось в рамках программы Harvard-Sussex Program в отделе исследований научной политики бизнес-школы Университета Сассекса. Финансирование проекта осуществлялось Центром по предотвращению распространения оружия и контролю над вооружениями Министерства иностранных дел и международного развития Великобритании. Мы благодарим Шонну МакИвор и Боаза Чана, оказавших неоценимую помощь в работе над проектом.

Руководитель исследования:
Д-р Джошуа Р. Мун (Dr. Joshua R Moon)
Эл. почта: J.R.Moon@sussex.ac.uk

Как лучше управлять ИИ в контексте ХБО?

1. Контекстуализировать разработку и применение ИИ для лучшего понимания проблем, связанных с конкретной политикой.
2. Социализировать ИИ путем определения участников разработки и целей использования.
3. Изучить проблемы ИИ, чтобы понять, как они могут ослаблять нынешние меры по запрещению ХБО или препятствовать их реализации.
4. Провести оценку существующих субъектов и архитектур систем управления, выявить пути решения проблем, связанных с ИИ, и, по возможности, устранить недостатки.
5. Сконцентрировать внимание на существующих инструментах и механизмах решения проблем, связанных с использованием ИИ.

Следующие шаги для улучшения политики в отношении ИИ:

- Построение более подробных потенциальных сценариев воздействия ИИ на запрещение ХБО с детальным описанием использования технологий и роли различных субъектов в рамках этих сценариев позволит лучше понять, как можно снизить риски;
- Понимание характера и последствий воздействия ИИ в контексте и на практике позволит более четко понять, как можно управлять ИИ, сохраняя его положительное влияние на общество;
- Анализ и оценка существующего управления режимами запрещения ХБО с опорой на подробные сценарии даст разработчикам политики более полное представление о точках взаимодействия ИИ с различными системами предотвращения и принципах создания таких систем с учетом этой новой области науки и техники.