Анализ инициативы Google DeepMind по обеспечению биорезистентности с помощью ИИ

Источник: AI News Healthcare

Оригинал: https://www.artificialintelligence-news.com/news/examining-google-deepmind-ai-bioresilience-push/

Gemini | биобезопасность | регулирование | риски ИИ | этика ИИ

Google DeepMind и Isomorphic Labs представили программу биорезильентности (биологической устойчивости), направленную на предотвращение неправомерного использования ИИ в биологии и одновременную помощь в реагировании на вспышки заболеваний.

Обе организации опубликовали обновленную информацию о совместной инициативе, которая началась незаметно, а за последние 12 месяцев переросла в более чем 15 партнерств с государственными органами, организациями по биобезопасности и исследовательскими группами.

Это раскрытие информации сопровождается специфической проблемой формулировок. Передовые модели (Frontier models), такие как Gemini, обладают все более глубокими знаниями в области биологии, и DeepMind признает, что сочетание этих систем со специализированными биологическими моделями, агентами (например, платформой Antigravity) и сторонними базами данных будет только усиливать эти возможности.

Однако те же знания, которые помогают исследователю определить мишень для вакцины, в теории могут помочь злоумышленнику устранить пробелы в собственных знаниях. DeepMind и Isomorphic описывают это как «двойной мандат»: способствовать научному прогрессу, который делает возможным передовой ИИ, и в то же время не допускать попадания этих же инструментов в руки людей, которые могут использовать их во вред.

Согласно заявлениям компаний, программа опирается на три столпа: предотвращение злоупотреблений, более быстрое обнаружение вспышек и реагирование на уже начавшуюся вспышку или атаку.

Более 15 партнерств, созданных за последний год, затрагивают все три направления, хотя в обновлении приводится ограниченная информация о том, какие именно организации участвуют, за исключением нескольких названных коллаборатороов, включая Ливерморскую национальную лабораторию имени Лоуренса, Институт безопасности ИИ Великобритании (UK AI Security Institute), CEPI (Коалиция за инновации в области готовности к эпидемиям) и Институт Фрэнсиса Крика.

DeepMind заявляет о намерении расширить эти связи в течение следующих шести-двенадцати месяцев, сосредоточив внимание на анализе угроз (threat intelligence), методах оценки ИИ-агентов и способах противодействия взлому (jailbreak mitigations). Компания также координирует свои действия с Форумом передовых моделей (Frontier Model Forum) по таким вопросам, как обработка более рискованных категорий обучающих данных, примером которых служат наборы данных по вирусологии.

Ограничение Gemini без блокировки легитимной науки

Работа по предотвращению основывается на моделировании угроз, предназначенном для выявления того, какие субъекты с наибольшей вероятностью попытаются совершить злоупотребление и какие препятствия мешают им в настоящее время. DeepMind сообщает, что использует сочетание экспертного «красного тестирования» (red-teaming) и рандомизированных контролируемых испытаний, чтобы оценить, может ли Gemini помочь кому-либо преодолеть эти препятствия.

Методы обучения после основного этапа (post-training) призваны научить модель отклонять вредоносные запросы, избегая при этом того, что компания называет «избыточным отказом» (over-refusal) на легитимные научные вопросы — баланс, который оказался сложным для всей отрасли в целом, а не только для DeepMind. Для выявления рискованной деятельности в режиме реального времени развертываются классификаторы и зонды (probes), а компания также проводит целевой анализ логов, чтобы обнаружить более тонкие паттерны злоупотреблений, которые могут пропустить автоматизированные фильтры.

Ни один из этих методов защиты не считается окончательно решенным. DeepMind представляет их как непрерывный процесс, а не как завершенную систему, что важно для любого предприятия или государственного органа, оценивающего, можно ли полагаться на текущие конфигурации защитных механизмов. Классификатор, настроенный на известные паттерны взлома в ходе контролируемой оценки, не гарантирует эквивалентной эффективности против новых методов атак, возникающих при реальном использовании, и компания не утверждает обратного.

Проблема скрининга синтеза ДНК

Один из более конкретных рисков, находящихся в стадии изучения, связан с синтезом ДНК. Компании, входящие в Международный консорциум по синтезу генов (International Gene Synthesis Consortium), в настоящее время проверяют заказы на соответствие спискам известных опасных патогенов и токсинов, используя алгоритмы скрининга. DeepMind прямо заявляет, что этот подход начинает давать сбои, поскольку ИИ теперь может помочь в проектировании последовательностей ДНК с функциями, аналогичными опасным патогенам, при этом их последовательность не будет достаточно близка к известным, чтобы сработали существующие системы скрининга.

Предлагаемое решение заимствует элементы существующей системы цифровых водяных знаков DeepMind — SynthID, которая, по словам компании, стала отраслевым стандартом для маркировки изображений и текстов, созданных ИИ. Адаптация этой системы к биологическим последовательностям представлена как исследовательская работа, а не готовый продукт.

Долгосрочная цель, описываемая скорее как открытая техническая задача, чем нечто близкое к решению, заключается в создании скрининга, который предсказывает, будет ли новая последовательность ДНК токсичной или патогенной на основе её функции, независимо от того, напоминает ли она что-либо в существующих базах данных.

Дешевое секвенирование как уровень обнаружения

Обнаружение зависит от метагеномного секвенирования, которое характеризует каждый микроорганизм в образце, а не проверяет лишь краткий список известных патогенов, как это делают традиционные методы

диагностики. Ограничивающим фактором является стоимость, и масштабирование этого подхода в регионах, где вспышки наиболее вероятны, требует значительного снижения цены.

В качестве одного из шагов к этой цели DeepMind указывает на сотрудничество между Google и Pacific Biosciences, в котором использовался кодинг-агент AlphaEvolve для повышения точности секвенирования. Компания заявляет, что сейчас изучает дальнейшие возможности — от оптимизации алгоритмов обработки данных секвенирования до участия в проектировании аппаратного обеспечения — и отдельно исследует, может ли AlphaGenome помочь в характеристике патогенов непосредственно на основе данных секвенирования.

Это остается исследовательским сотрудничеством, а не системами, развернутыми в полевых условиях; дистанция между повышением точности секвенирования в контролируемом конвейере и функционирующей сетью раннего предупреждения в сточных водах и транспортных узлах в условиях ограниченных ресурсов весьма велика.

Рекорд публикаций AlphaFold и разрыв в контрмерах

Столп «реагирования» опирается на проблему дефицита медицинских контрмер, из-за которой многие известные патогены остаются без лицензированных средств диагностики, вакцин или методов лечения. DeepMind приводит более 10 000 публикаций по инфекционным заболеваниям, в которых за последние пять лет упоминался AlphaFold, включая работы по туберкулезу, передаче малярии и картированию мишеней для таких угроз, как mpox и вирус Нипах.

Новейшим дополнением к этому списку стало партнерство с программой биорезильентности Ливерморской национальной лаборатории, которая планирует использовать AlphaFold 3 для разработки антител широкого спектра действия, включая усилия по созданию пан-филовирусных антител. DeepMind заявляет, что в этом году продолжит добавлять структуры и комплексы белков в базу данных структур белков AlphaFold (AlphaFold Protein Structure Database), уделяя приоритетное внимание мишеням, важным для разработки контрмер.

Доступ к новым агентным системам, включая Co-Scientist, предоставляется избранным исследователям, среди которых ученые из национальных лабораторий Министерства энергетики США, работающие в рамках миссии Genesis.

Isomorphic Labs пошла еще дальше, создав специализированное подразделение, предназначенное для быстрого развертывания своего движка по разработке лекарств во время новой вспышки, работая совместно с правительственными и национальными исследовательскими органами, такими как Ливерморская лаборатория, Институт безопасности ИИ Великобритании, CEPI и Институт Фрэнсиса Крика. Компания также пообещала выделить 7 миллионов долларов программе Health for Human Potential (программа Philanthropy Asia Alliance) для исследований инфекционных заболеваний в Азии.

Рекомендации DeepMind для политиков США напрямую соответствуют её трем столпам и опираются на конкретные законопроекты, находящиеся на рассмотрении:

- **В области предотвращения:** компания поддерживает федеральную структуру безопасности передового ИИ, закон AI-Ready Bio-Data Standards Act (H.R. 7907), обязательный скрининг синтеза ДНК через закон Biosecurity Modernization and Innovation Act (S. 3741) и закон SCALE Biology Act (H.R. 8981).

- **В области обнаружения:** компания выступает за расширение метагеномного секвенирования в транспортных узлах и густонаселенных центрах при поддержке закона America’s Living Library Act (S. 4023) и дополнительного финансирования со стороны DARPA (Управление перспективных исследовательских проектов Министерства обороны США) и HHS (Министерство здравоохранения и социальных служб США) для исследований раннего предупреждения.

- **В области реагирования:** компания призывает к принятию закона Web of Biological Data Act (H.R. 9307 / S. 4770) и инвестициям в производственные мощности, которые поддерживались бы в «горячем» состоянии и были готовы к быстрой активации, наряду с заранее созданными сетями клинических испытаний и ускоренными регуляторными путями.

Ни один из этих законопроектов еще не принят, и именно в разрыве между политическими пожеланиями компании и функционирующей федеральной системой биобезопасности будет заключаться настоящая проверка этой программы в течение следующих 6–12 месяцев.