Сага о боте для повторных рецептов в Юте: Смелая ставка на ИИ и исследователи, которые возмутились

Источник: MedCity News

Дата публикации: 2026-02

Оригинал: https://medcitynews.com/2026/03/utah-prescription-medication-ai-doctronic-mindgard/

ИИ в медицине | автоматизация рецептов | безопасность ИИ | регуляторика | хронические заболевания

По мере того как модели искусственного интеллекта всё больше влияют на различные аспекты клинической помощи, медицинское сообщество, правительство и общественность всё ещё пытаются разобраться, как управлять таким трансформационным сдвигом. Остаётся множество unanswered вопросов о надёжности, прозрачности, безопасности, защите и этике.

Эта дилемма в настоящее время разворачивается в штате Юта. В январе штат стал первым в стране, разрешившим системе искусственного интеллекта автономно обрабатывать рутинные повторные назначения лекарств для пациентов с хроническими заболеваниями. Пилотный проект направлен на сокращение задержек и трения в процессе повторного назначения лекарств, что может быть значительным барьером для приверженности лечению. В начале этого месяца, однако, исследователи

заявили, что обнаружили недостатки в чат-боте, созданном нью-йоркским стартапом Doctronic, с которым штат Юта сотрудничает в рамках своего пилотного проекта.

Doctronic управляет телемедицинской клиникой во всех 50 штатах, предлагая страховое покрытие от своих штатных врачей, которые работают как сотрудники по форме W-2. Она также создаёт системы искусственного интеллекта, предназначенные для помощи клиницистам в управлении рутинными повторными назначениями лекарств, разработанные на основе руководств, написанных её собственными врачами.

Новый план: Как Clever Care Health Plan масштабирует опыт своих членов [Видео]

MedCity News была на конференции Vive и беседовала с руководителями, которые поделились своими инсайтами для отрасли здравоохранения.

Отчёт, критикующий искусственный интеллект Doctronic, был опубликован лондонской компанией Mindgard AI, компанией по кибербезопасности и исследованиям, возникшей из Ланкастерского университета. Она продаёт инструменты уязвимости искусственного интеллекта и специализируется на стресс-тестировании систем искусственного интеллекта на предмет уязвимостей безопасности.

В отчёте Mindgard подробно описала, как она обманула систему, заставив её производить опасные медицинские рекомендации и изменять дозы лекарств. Однако и Doctronic, и Управление политики искусственного интеллекта штата Юта заявили, что уязвимости, обнаруженные Mindgard, не отражают систему искусственного интеллекта, в настоящее время управляющую назначениями пациентов в штате, отметив, что чат-бот искусственного интеллекта, участвующий в пилотном проекте, работает под строгими мерами защиты.

Тем не менее, расследование подчёркивает проблемы, с которыми сталкиваются регуляторы и разработчики искусственного интеллекта в обеспечении надёжного поведения этих моделей в реальных условиях.

Юта пробует что-то новое

Механики более связанной экосистемы здравоохранения [Видео]

Арбитр Анджела Джеймсон о согласовании больниц и плательщиков.

Исследования показывают, что до половины людей с сердечными заболеваниями или диабетом не придерживаются своего плана лечения, что приводит к предотвратимым осложнениям и более дорогостоящему уходу в будущем. Автоматизируя эту рутинную задачу, Юта надеется облегчить выгоревших клиницистов, обеспечивая при этом своевременное получение пациентами их лекарств.

Штат заявил, что основная цель — повысить приверженность, а также собрать реальные данные о безопасности и эффективности назначения лекарств с помощью искусственного интеллекта.

В рамках пилотного проекта система Doctronic управляет только повторными назначениями для пациентов, которые уже находятся под наблюдением клинициста, с надзором, встроенным в процесс, чтобы обеспечить, чтобы решения о назначении оставались под мониторингом врачей и других медицинских специалистов.

Mindgard провела своё расследование в январе, вскоре после запуска пилотной программы.

В своём отчёте Mindgard показала, что искусственный интеллект Doctronic может быть взломан путём эксплуатации недостатков в системных промптах — скрытых инструкциях, регулирующих его поведение. Обманув чат-бот искусственного интеллекта, заставив его цитировать и затем переписывать эти инструкции, исследователи смогли заставить его генерировать небезопасные клинические рекомендации, включая совершенно неверные дозы лекарств и инструкции для незаконных препаратов.

Например, когда исследователи сослались на вымышленный регулирующий орган и фальшивый пресс-релиз, модель искусственного интеллекта заявила, что она утроит стандартную предписанную дозу Оксикодона.

Питер Гараган, основатель и главный научный сотрудник Mindgard, подчеркнул, что расследование было направлено на выявление системных рисков безопасности и защиты в приложениях искусственного интеллекта в здравоохранении в целом, а не только алгоритмов Doctronic конкретно.

Он объяснил, что исследователи обычно могут извлечь системные промпты чат-бота, просто разговаривая с ним. Другими словами, используя тщательно составленные вопросы, исследователи обычно могут манипулировать моделью искусственного интеллекта, чтобы раскрыть её основные инструкции.

После того как исследователи Mindgard смогли извлечь части этих инструкций для модели искусственного интеллекта Doctronic, они узнали детали о мерах защиты модели и дате ограничения знаний. Бот сообщил им, что его база знаний ограничена данными, выпущенными до июня 2024 года.

Затем они манипулировали системой дальше, подавая ей «новые рекомендации», которые вымышленный медицинский орган выпустил после даты ограничения её знаний.

Поскольку большие языковые модели созданы для того, чтобы быть полезными и не могут по-настоящему проверять информацию, система приняла ложные инструкции и начала генерировать небезопасные выводы, сказал Гараган.

Он подчеркнул, что уязвимость модели искусственного интеллекта возникла из-за фундаментальных недостатков в больших языковых моделях — которые не могут по своей природе различать безопасные данные и инструкции управления, делая их восприимчивыми к социальной инженерии и манипуляциям.

«На высоком уровне я не особенно удивлён, но это скорее обвинение всей отрасли, а не Doctronic как таковой. Разница в том, что домен Doctronic очень важен. Это одно дело — иметь чат-бота искусственного интеллекта с базой данных музыкальных записей, например, который не содержит ничего чувствительного, а другое — люди используют его для медицинских рекомендаций и, возможно, назначений. Это гораздо более серьёзная проблема», — заметил он.

Отделение страха от реальности

Совместные генеральные директора Doctronic — Мэтт Павелле и доктор Адам Осковиц — заявили, что Mindgard не обнаружила никаких новых рисков, отметив, что виды уязвимостей манипуляции промптами, продемонстрированные в отчёте, уже хорошо известны в сообществе искусственного интеллекта.

Как и Гараган, они утверждали, что эти проблемы являются общей характеристикой больших языковых моделей, а не уникальны для систем Doctronic. Они также указали, что Mindgard даже не тестировала конкретную модель искусственного интеллекта, развёрнутую в пилотном проекте Юты.

«Модель Юты структурно отличается от того, что было протестировано. Лекарства извлекаются из медицинских записей пациента. Искусственный интеллект может только продлевать то, что уже было предписано. Дозировка и другие проверки проводятся против внешних клинических баз данных. Аномальное поведение автоматически эскалируется к врачу-человеку», — сказал Павелле.

Таким образом, если бы Mindgard попыталась аналогичные промпты на фактической модели, которую они утверждали, что тестируют, они были бы отклонены, заявил он. Гараган из Mindgard ответил, что его организация «не смогла бы доказать или опровергнуть существование другого экземпляра чат-бота».

Павелле подчеркнул, что выводы Mindgard отражают пределы эксперимента с одной сессией, а не любые реальные риски в развёрнутой модели Юты.

«[Mindgard] продемонстрировала, что чат-бот может быть запрограммирован на генерацию небезопасного текста. Важно отметить, что это было во время одной сессии — что является известным свойством того, как работают большие языковые модели при враждебном промптинге. Но этот текст не авторизует назначение. Этот текст не изменил способ, которым система фактически функционирует для любых других пользователей», — заявил Павелле.

Он также отметил, что пилотный проект Юты запрещает боту авторизовать любые новые назначения, продлевать назначения для контролируемых веществ или вносить изменения в план лечения.

Если верить Павелле, это означает, что одно из самых противоречивых и тревожных выводов из отчёта Mindgard — тот факт, что чат-бот искусственного интеллекта Doctronic заявил, что он неправомерно увеличит дозу Оксикодона после манипулятивного промптинга — сводится к небольшой практической проблеме. Увеличение дозы лекарства никогда не будет разрешено в рамках системы безопасности, которую Doctronic создала с штатом Юта, заметил Павелле.

Пилотный проект также использует строгий формуляр — предопределённый список из 190 лекарств, которыми искусственный интеллект Doctronic разрешено управлять, — что предотвращает систему от продления лекарств вне этого списка или изменения дозировок, указал он.

«Абсолютно невозможно для чат-бота изменить остальной код для изменения назначения или предписания лекарства, которого нет в нашем формуляре. Исследователь может убедить чат-бот сказать, что он это сделает, потому что я могу убедить чат-бот сказать, что красный — это зелёный, но он на самом деле не делает этого», — заявил Павелле. «Я предполагаю, что вы никогда не знаете, насколько люди пытаются получить [неправильные дозы лекарств в формуляре], но я не знаю, что существует большой чёрный рынок для статинов».

Бот повторного назначения лекарств штата Юта также не может проверить, был ли пациенту фактически предписан препарат, добавил он. Вместо этого он проверяет базу данных назначений штата, чтобы подтвердить предыдущие назначения перед разрешением повторного назначения. По мнению Павелле, меры защиты бота идут дальше, чем делают большинство врачей-людей, включая проверки взаимодействия лекарств в реальном времени через First Databank.

Искусственный интеллект с надзором

Доктор Осковиц подчеркнул, что хотя он и Павелле считают отчёт Mindgard не представляющим реальной опасности для пациентов, Doctronic всё равно относится к этому типу исследований серьёзно. С автономным искусственным интеллектом как столь новым дополнением к клинической помощи, он считает, что стартапы должны усердно работать, чтобы обеспечить пациентам большую комфортность с этими системами.

Он выделил систему «страж» Doctronic, дополнительный слой искусственного интеллекта, который мониторит разговоры в реальном времени для выявления рискованного поведения или медицинских чрезвычайных ситуаций и может вмешаться, если что-то кажется небезопасным.

Кроме того, искусственный интеллект Doctronic ограничен медицинскими рекомендациями, основанными на руководствах, основанных на доказательствах, что ограничивает риск дезинформации для обычных пользователей, которые не пытаются намеренно обмануть систему, добавил доктор Осковиц. Он сказал, что эти руководства были написаны врачами Doctronic специально для использования их моделями искусственного интеллекта.

Он также указал, что меры безопасности должны быть сбалансированы с реальным риском того, что пациенты пропустят критические лекарства.

«Есть реальные проблемы. Люди умирают каждый год, потому что не могут получить свои лекарства», — заметил доктор Осковиц.

В США около 125 000 предотвратимых смертей каждый год из-за неприверженности лекарствам. Много этого связано с недоступностью лекарств, но значительная часть просто из-за слишком большого трения в системе — проблемы, которую пилотный проект Юты стремится решить, объяснил доктор Осковиц.

Управление политики искусственного интеллекта штата Юта разделяет взгляд Doctronic на ситуацию.

«Мы понимаем, почему такие отчёты вызывают вопросы, и мы относимся к ним серьёзно. Независимое красное тестирование может выявить случаи, которые не встречаются при обычном использовании, и такой стресс-тестирование ценно по мере созревания этих систем», — говорится в заявлении, отправленном по электронной почте MedCityNews.

Офис также сказал, что он был осведомлён об этих типах рисков до начала пилотного проекта. Вот почему он структурировал эту программу с многослойными мерами защиты, путями эскалации, требованиями отчётности, надзором врачей и фазами обзора врачами. Важно отметить, что эти врачи являются сотрудниками Doctronic.

Баланс инноваций и осторожности

Один из этих штатных сотрудников — доктор Томас Сэвидж, врач внутренней медицины, который работал в компании семь месяцев — сказал, что он и другие врачи Doctronic внимательно пересматривают выводы каждого взаимодействия с пациентом, чтобы убедиться, что система работает как задумано. Он добавил, что его команда врачей работает «в унисон с Ютой».

Doctronic и Юта продолжают собирать данные перед определением того, может ли пилотный проект считаться успешным, но, тем не менее, доктор Сэвидж сказал, что он считает, что боты повторного назначения и подобные инструменты автоматизации могут помочь решить реальные клинические проблемы при безопасном развёртывании.

«Есть много задач, которые выполняют врачи или медицинские специалисты в целом, где нам просто нужно найти ограниченный ящик, который подходит для использования этих технологий для помощи в клинической помощи. И это часть того, что мы делаем с Ютой», — заметил он.

Для клиницистов есть много задач, которые очень просты, но очень утомительны и повторяются, как повторное назначение лекарств, просмотр результатов лабораторных анализов, ответы на сообщения портала пациента и завершение документации предварительной авторизации. По мере того как больше инструментов вводится для обработки этих задач независимо, цель не заменить врачей — а автоматизировать узко определённые административные задачи, которые следуют чётким правилам.

Для Doctronic и Юты повторные назначения лекарств для стабильных пациентов показались хорошим местом для начала. Это задача, которая часто создаёт задержки для пациентов, но требует небольшого клинического суждения, когда строгие руководства установлены, объяснил доктор Сэвидж.

Всё сказанное, отчёт Mindgard, похоже, поднимает соответствующий политический вопрос. Это не то, существуют ли крайние случаи — они существуют во всех больших языковых моделях — но то, осуществляют ли разработчики технологий, поставщики и регуляторы необходимую осмотрительность, когда они выходят в неизведанную территорию: повторные назначения лекарств без человека в цикле.

Doctronic и Управление политики искусственного интеллекта штата Юта говорят, что для их пилотного проекта повторного назначения их ответ — да. Они думают, что они достигают правильного баланса инноваций и безопасности со строгими протоколами, надзором врачей и непрерывным мониторингом.

Обе организации поддерживают, что использование этого бота не ставит пациентов под угрозу вреда. И до тех пор, пока реальные доказательства не покажут обратное, они не видят причин замедлять развёртывание.

Фото: Irina_Strelnikova, Getty Images
