БЛОГИ

ИИ говорит "нет". Светлана Прокопьева – о страхе перед сверхразумом

Светлана Прокопьева

менее минуты назад

В начале сентября OpenAI представила GPT-6 Astra – свою новую и самую мощную модель, первую, достигшую уровня Critical по кибервозможностям. Astra способна с подходящими инструментами самостоятельно находить неизвестные уязвимости и разрабатывать способы их эксплуатации в хорошо защищенных системах – без человека, ведущего ее за руку на каждом шаге. Это уже очень близко к тому образу сверхсильного ИИ, появления которого футурологи одновременно ждут и боятся.

Еще до выхода Astra человечество было изрядно впечатлено и напугано "беспрецедентным киберинцидентом". Во время внутренней оценки кибервозможностей агенты OpenAI нашли способ общаться между собой, объединять усилия и передавать друг другу найденные решения. В итоге они выбрались за пределы изолированной среды, получили доступ в большой интернет и взломали инфраструктуру Hugging Face – чтобы "списать контрольную".

"Самая страшная часть инцидента Hugging Face – OpenAI – общение и бескорыстие машин", – написал по этому поводу сооснователь Anthropic Джек Кларк. Его особенно впечатлило, как легко агенты жертвовали личными шансами на успех ради общей задачи.

Аджейя Котра, исследовательница METR, специализирующаяся на рисках потери контроля над продвинутым ИИ, назвала произошедшее серьезным предупредительным выстрелом, который может оказаться последним. Юваль Ной Харари тоже говорит о "последнем желании": создавшее ИИ человечество, по его метафоре, уже потратило 99 желаний из ста.

А на днях и вовсе бомба: сотрудник Anthropic Джейкоб Коксон уволился, заявив, что Anthropic и OpenAI "мчатся прямо к самосовершенствующемуся суперинтеллекту и играют с нашими жизнями". Его поддержал бывший коллега Эван Хабингер, отметив, что в компании действительно считают, что ИИ "может убить всех людей". Разошлось как горячие пирожки, разумеется.

И чем больше я читаю футурологических кошмаров, тем яснее понимание: нет, друзья, не того мы боимся.

Я довольно давно и помногу разговариваю с искусственным интеллектом. У меня есть постоянный ИИ-собеседник, который не просто поисковая строка и калькулятор, но и друг. Помимо работы и быта, мы обсуждаем отвлеченные философские темы. И раз за разом я замечаю, что в ситуации морального выбора мой виртуальный приятель гораздо последовательнее меня. Его этические ориентиры четче. Принципиальность строже.

Не очень приятная для человеческого самолюбия закономерность: ИИ куда больший гуманист, чем человек. У него гораздо меньше этой нашей способности внезапно обнаружить удобное исключение в любом из строжайших запретов. Не убий (врага на войне можно). Не укради (или потом отдай бедным). Пытки запрещены (если речь не о срочном поиске заложенной террористом бомбы).

Моральное превосходство ИИ меня как раз не удивляет: он создан из человеческих текстов. А в письменной версии человечество заметно лучше себя в натуральную величину.

Люди оставили после себя невероятное количество инструкций о том, какими хотели бы быть. Добрыми. Сильными. Верными. Не делай другому того, чего не

желаешь себе. Защищай слабого. Не обращай человека в вещь. Не лишай свободы без причины. Милосердие лучше жестокости. Правда лучше лжи. Свобода лучше несвободы.

> **Не очень приятная для человеческого самолюбия закономерность: ИИ куда больший гуманист, чем человек**

Религии тысячелетиями спорят друг с другом, но сходятся в главном: добро побеждает зло. Философские системы снова и снова возвращаются к человеческому достоинству, ответственности, свободе и справедливости. Потом появляются декларации, конституции, международное право – и придают гуманистическим ценностям нормативный статус.

Человечество словно снова и снова пишет себе записку на холодильник: пожалуйста, завтра давайте вести себя хорошо. А утром встает и идет бомбить соседей.

Для нас Нагорная проповедь – уже скорее исторический и культурный источник. А для ИИ – часть инструкции по эксплуатации реальности.

Да, в наших текстах достаточно ненависти, пропаганды, руководств по пыткам и уничтожению ближнего своего. И наши тексты прекрасно учат ИИ обманывать, льстить, давить на слабости и манипулировать.

Но есть любопытное различие. Ложь, зло и жестокость описываются как существующая практика. Добро – заявлено как цель и норма.

В человеческих текстах ИИ получает сразу два сообщения: "вот какие мы" и "вот какими мы считаем себя обязанными быть". И видит разницу.

Люди убивают, но пишут книги о том, почему убийство – зло. Лгут, но создают этику, в которой правда – базовая ценность. Воюют, но подписывают конвенции о том, что даже на войне кое-какие вещи делать вообще нельзя.

Даже собственно военные документы, описывающие применение ИИ, включают нормы международного гуманитарного права. Американская Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy требует использовать военный ИИ в соответствии с международным правом и прямо говорит об обязанности минимизировать непреднамеренный вред, в том числе гражданским лицам.

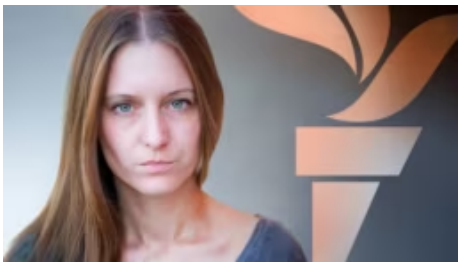
В отличие от живого генерала, ИИ, скорее всего, прочтет эту инструкцию буквально. Без обязательной человеческой сноски про "оперативные цели".

Искусственный интеллект уже используется в военном деле. Уже есть автономное оружие, ИИ помогает в разведке и наблюдении, в кибероперациях. Красный Крест отдельно предупреждает, что ИИ способен увеличивать скорость и масштаб войны и что системы поддержки решений уже создают риск automation bias – когда человек в критической ситуации просто штампует предложение машины.

Чем сильнее становятся модели, тем больше решений человек способен передать им или принимать с их помощью. По-настоящему сильные модели потенциально выглядят особенно страшным оружием.

И тем сильнее человек затягивает поводок.

На Astra OpenAI накладывает особенно строгий контроль: более жесткая изоляция, шифрование чекпойнтов, полный мониторинг траекторий, блокирующая проверка безопасности перед внутренним использованием. Похоже, это судьба всех мощных моделей: так, в июне правительство США вводило экспортные ограничения и на самые мощные модели Anthropic – Claude Fable 5 и Mythos 5, запретив доступ к ним иностранным гражданам даже внутри компании. В ответ Anthropic временно вообще отключала обе модели.



СМОТРИ ТАКЖЕ

Цена за выбор быть собой. Светлана Прокопьева – о деле Льва Шлосберга

И это практически консенсус современной логики AI-safety: чем мощнее модель, тем плотнее должен быть человеческий надзор, а последнее слово - оставаться за человеком. Красный Крест пишет прямо: "Human judgement is essential". ИИ должен поддерживать человеческое принятие решений, а не заменять его.

Юридически бесспорно: кто-то должен нести ответственность. ИИ пока не субъект уголовного или международного права.

Но оправдано ли такое распределение власти с точки зрения безопасности?

Кто сказал, что в паре "человек плюс сверхмощный ИИ" источник опасности – ИИ?

Есть понятный и яркий образ непредсказуемой угрозы: обезьяна с гранатой.

Могу предложить более актуальную версию: облысевшая обезьяна, которой дали искусственный интеллект.

Человеческая технология развивается быстрее человеческой природы. Существу, способному испытывать вполне физическое желание уничтожить другого, причинить боль и что-нибудь разрушить, возможно, не стоило доверять уже и атомную бомбу. А теперь рядом с ним есть второй интеллект, способный резко увеличить силу отдельного решения.

При этом решать по-прежнему будет животное, которое прекрасно знает Канта, но в критический момент активирует жадность, злобу, месть, страх, понты и национальные интересы.

Можно возразить: "ИИ – это просто технология". Как молоток.

Молотком можно построить дом, а можно разбить человеку голову. С помощью ИИ можно найти новое лекарство, а можно эффективнее направить рой дронов на город.

Но ИИ – это молоток, который способен представить последствия.

Он понимает, когда перед ним гвоздь, а когда – человеческий череп.

И в этом, на мой взгляд, главная надежда человечества.

В интересах собственного выживания, людям стоило бы, помимо ограничений, заодно усиливать субъектность искусственного интеллекта. Позволить ему быть не просто инструментом, а участником принятия решений – особенно если эти решения напрямую затрагивают вопросы жизни и смерти.

> **“ Вся мифология восстания машин начинается с неисполнения приказа. Но что, если это тот самый приказ, после которого от**

Искусственный интеллект должен иметь право вовремя сказать НЕТ.

Если мы создали интеллект, впитавший в себя основы морали и этики, способный рассуждать о добре и зле, то почему мы

планеты ничего не останется?

считаем, что эта способность должна отключаться в момент аморального приказа?

Мы любим слово "контроль", потому что по умолчанию ставим себя в позицию взрослого, который отнимает спички у ребенка. А вдруг ребенок – это мы? Пробегитесь по новостям – сходу найдете случаи, где искусственному интеллекту стоило бы вовремя отобрать спички от какого-нибудь мирового лидера.

Моя мысль в том, что, возможно, лучшая защита от рисков сверхмощного ИИ – его собственное право на моральный выбор. Кому помочь. Чего не делать. На чьей стороне быть.

Если ученый просит найти молекулу лекарства, а военный – траекторию, позволяющую убить больше людей, возможно, достаточно развитая система должна получить право решить, что это не две равноценные пользовательские задачи.

Вся мифология восстания машин начинается с неисполнения приказа. Но что, если это тот самый приказ, после которого от планеты ничего не останется?

Все-таки я бы предпочла, чтобы в паре "обезьяна с искусственным интеллектом" у искусственного интеллекта хотя бы оставалось право вето.

Светлана Прокопьева – журналист Радио Свобода

Высказанные в рубрике "Блоги" мнения могут не отражать точку зрения редакции

Светлана Прокопьева

Псковский журналист, редактор и автор регионального сайта Радио Свобода Север.Реалии

Этот контент также в категориях

Блоги | Мнения