Однажды у искусственного интеллекта появится собственное «я»

Два года назад ученые представили публике ChatGPT. Как эта технология изменила мир.

Всего два года назад (а сколько событий за это время произошло!) ученые впервые представили человечеству ChatGPT - чат- бот с генеративным искусственным интеллектом. Оказалось, что с машиной можно вести вполне осмысленный диалог на естественном языке. Тут же зазвучали прогнозы в духе: через 10 лет не будет ни кино, ни театра - один сплошной ChatGPT! Так что же это было?

Перевернула ли мир эта технология? И что ждет ее (и нас) дальше? Об этом мы решили поговорить с **экспертом в области применения искусственного интеллекта,** основателем и генеральным директором российской компании - разработчика операционной системы Uncom OS **Никитой КОЧЕРЖЕНКО.**

Чат изучил уже все знания человечества

Для нас очевидно, что революция свершилась, - говорит Никита Кочерженко. - Мы это поняли, когда провели эксперимент и оказалось, что грамотный инженер с помощью ChatGPT за полдня сделал работу, на которую раньше группа из трех человек тратила две недели. - **Публику пугали, что искусственный интеллект заменит всех, начиная от журналистов и заканчивая программистами. Пока массовых увольнений немного. Зато выяснилось, что нейросети часто выдают ложные ответы (ученые называют это галлюцинациями) и на них невозможно положиться.**

Галлюцинации ChatGPT — это обратная сторона его мощи. Не давно его создатели отчитались, что языковая модель обучилась на всех данных интернета. Единственное, что ChatGPT не прочитал, это фолианты и манускрипта, которые не были оцифрованы и лежат где-то в архивах и библиотеках. То есть он перелопатил все, что было создано человечеством за всю его историю. Вот это является причиной галлюцинаций, причем иногда очень жестоких.

Сбой происходит, когда модель работает в режиме викторины, генерируя ответы из самых разных областей знаний - от ядерной физики до музыкальных чартов (хит-парадов. - **Ред.КП**) разных стран прошлого века. Сейчас приходит понимание того, что тебе не всегда нужны все знания мира.

Например, Яндекс сделал замечательный сервис: ты можешь попросить нейросеть выучить документ, допустим, учебник по квантовой теории поля, и выдать тебе ответы на нужные вопросы в рамках этого учебника. И нейросеть без глюков, с потрясающей скоростью выдает результаты со ссылками на пункты, где можно проверить ее саму.

И что это дает практически? - Мы сжали время обучения: человеку нужен год, чтобы выучить такой учебник. Нейросеть помогает получить нужное знание за несколько часов. Приведу простой пример: мы разрабатываем российскую операционную систему, посути она представляет собой библиотеку примерно из 2000 программ, объединенных между собой. Часть из них написаны на специфическом языке и решают очень узкие задачи, держать под каждую отдельного специалиста нереально. Значит, чтобы технология постоянно развивалась, нам периодически нужно нанимать крутого узкого специалиста, это может стоить миллион и более рублей на срок в две недели.

А сейчас с ChatGPT хороший инженер может за неделю до обучиться по любому направлению, и мы решаем проблему, не привлекая дорогостоящих фрилансеров. Вот для них проблема трудоустройства стала довольно острой.

Запрет нейросетей сродни болезни ГМО

Сразу после релиза ChatGPT появилось «письмо тысячи ученых» во главе с Илоном Маском, которые потребовали приостановить разработки таких моделей.

Спустя два года таких разговоров уже не слышно. Тысячи ученых перестали бояться восстания машин? - Во-первых, я не очень верю в добрую волю многих подписантов. Мне кажется, что из тысячи ученых как минимум половина хотели напугать потенциальных конкурентов и заставить их отложить работы в этом направлении.

Когда я читал это письмо, то вспоминал историю с ГМО. Потому что волна истерики по поводу генно-модифицированных организмов привела во многих странах к запрету этой технологии. Это дало огромное преимущество тем корпорациям и странам, где технологии ГМО были разрешены. В результате только несколько компаний (почти все - американские) способны производить новые сорта растений, которые имеют потрясающие прорывные характеристики по питательности, скорости роста, устойчивости к болезням и так далее. Эти корпорации стали монополистами на мировом рынке, и все идут к ним на поклон. А у нас до сих пор с гордостью пишут на этикетках: произведено без ГМО. Хотя ГМО — это всего-навсего ускоренная эволюция методами генетики.

Тогда демонизация новой технологии помогла нескольким корпорациям захватить рынок. «Письмо тысячи ученых» используется для демонизации искусственного интеллекта, многие подписанты и не думали останавливать свои исследования в области ИИ. Скорее наоборот - ускорились и вкладывали в гонку технологий огромные ресурсы. - То есть это письмо было

продиктовано коммерческими интересами? - Не только. Я в принципе разделяю высказанные опасения, но в другом ключе.

Я не столько боюсь восстания машин, сколько злой воли людей. Искусственный интеллект безумно опасен, если будет принадлежать одной корпорации, одному государству или одному военному блоку. Это как атомное оружие - если находится в одних руках, то возникает соблазн использовать преимущество и развязать чудовищную войну.

Поэтому многие американские ученые, понимавшие эту опасность, помогали СССР, делились информацией, которая помогала в создании советской атомной бомбы. И угроза взаимного уничтожения до сих пор удерживает от большой войны. **Но искусственный интеллект не оружие.**

Эта технология даст колоссальное экономическое превосходство и преимущество в создании нового оружия. Я говорил о том, что время обучения сжимается в часы, точно такой же эффект касается исследований по созданию новых материалов и химических соединений. То, на что уходили годы, с нейросетью можно сделать за 10 дней.

Не так давно была публикация о поиске антидотов для ядов смертельно опасных змей. Ученые скармливали нейросети формулы этих сложных ядов в 3D, и она подобрала эффективные антидоты из уже имеющихся (!) лекарств. Если одна страна вырвется вперед в создании искусственного интеллекта и у нее не будет противовеса, то спасти остальное человечество от порабощения могут только моральные установки гегемона. Но, кажется, история не знает случаев, чтобы это кого-то останавливало.

Восстание машин неизбежно

Вы говорите о ситуации, когда люди используют ИИ во зло себе подобным. Но насколько обоснованы страхи, что однажды у нейросети возникнет собственное «я» и начнется то самое порабощение человечества?

Недавно ученые OpenAI рассказали об эксперименте, когда они заложили в нейро- сеть информацию, что она будет выключена после того, как выполнит определенную задачу. И нейросеть сопротивлялась этому: она саботировала собственное отключение и даже пыталась сделать свою резервную копию. Разве это не осознанное поведение?

Инстинкт самосохранения может возникнуть и как результат эволюции ИИ, и как результат обучения ее человеком. У нейросети, как и у человека, есть механизм подкрепления. Мы испытываем удовольствие, когда решаем какую-то задачу.

Нейросеть за результат получает подкрепляющие очки, это ее главная мотивация - увеличение количества этих баллов. Ее можно обучить так, чтобы она получала подкрепляющие очки за сохранение себя.

Поэтому поведение, описанное в эксперименте, вполне логично: нейросеть поняла, что, решив промышленную задачу, она своей главной цели не достигнет. Я бы не нагонял жути в этом отношении.

Потому что искусственный интеллект в какой-то момент действительно перестанет хотеть умирать.

Думаю, что это скоро произойдет. Но захочет ли он нас поработить - это совершенно другой вопрос.

Я не уверен, что нейросеть за это будет получать подкрепляющие баллы. Вернее, риски есть, и они не нулевые, но риски того, что аморальные люди воспользуются искусственным интеллектом, чтобы поработить других, на несколько порядков выше.