МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
УЧРЕЖДЕНИЕ НАУКИ
ВОЛОГОДСКИЙ НАУЧНЫЙ ЦЕНТР РАН

ИНФОРМАЦИОННЫЙ ВЫПУСК № 71 (2432)

Серия

«ЭКОНОМИЧЕСКИЕ ПРОЦЕССЫ»

ВолНЦ РАН продолжает знакомить своих подписчиков с наиболее интересными, на наш взгляд, публикациями, затрагивающими актуальные вопросы российской экономики и политики.

В выпуске представлена статья К. Меца «Какую точно опасность представляет искусственный интеллект?», опубликованная в журнале «Экономист», № 4, 2024 г.

Вологда
май 2024

Какую точно опасность представляет искусственный интеллект?

В конце марта более 1000 технологических лидеров, исследователей и других экспертов, работающих в сфере искусственного интеллекта, подписали открытое письмо, в котором предупреждают, что технологии искусственного интеллекта (ИИ) представляют «серьезные риски для общества и человечества».

Группа, в которую вошел Илон Маск, исполнительный директор Tesla и владелец глобальной социальной сети, призвала лаборатории искусственного интеллекта приостановить разработку своих самых мощных систем на шесть месяцев, чтобы можно было лучше понять опасности, скрывающиеся за этой технологией.

«Мощные системы искусственного интеллекта следует разрабатывать только в том случае, если мы уверены, что их эффект будет положительным, а риски – управляемыми», – говорится в письме.

Письмо, которое набрало более 27 000 подписей, было кратким. Его язык был широким. И некоторые из имен, стоящих за письмом, похоже, имеют противоречивые отношения с ИИ-технологиями. Например, И. Маск создает свой собственный стартап в области ИИ и при этом является одним из основных спонсоров организации, написавшей письмо.

Письмо отражает растущую обеспокоенность экспертов по искусственному интеллекту в связи с тем, что новейшие системы, в первую очередь GPT-4 – технология, представленная стартапом OpenAI из Сан-Франциско, – могут нанести вред обществу. Эксперты полагают, что будущие системы станут еще более опасными.

Некоторые риски уже появились. Другие не будут проявляться в течение месяцев или лет. Третьи являются чисто гипотетическими.

«Наша способность понять, что может пойти не так с очень мощными системами искусственного интеллекта, крайне слаба». – сказал Йошуа Бенджио, профессор и исследователь искусственного интеллекта в Университете Монреаля, – «Поэтому нам нужно быть очень осторожными».

Почему они обеспокоены?

Доктор Бенджио, пожалуй, самый важный человек, подписавший письмо. Работая с двумя другими учеными – Джеффри Хинтоном, до недавнего времени исследователем в Google, и Яном Ле-Куном, ныне главным научным сотрудником по искусственному интеллекту в глобальной социальной сети, – доктор Бенджио провел последние четыре десятилетия, разрабатывая технологию, которая управляет такими системами, как GPT-4. В 2018 г. исследователи получили премию Тьюринга, часто называемую Нобелевской премией в области техники вычислений, за свою работу над нейронными сетями.

Нейронная сеть – это математическая система, которая обучается навыкам путем анализа данных. Около пяти лет назад компании, как Google, Microsoft и OpcnAI, начали создавать нейронные сети, которые обучаются на огромных объемах цифрового текста. Нейронные сети называют большими языковыми моделями, или LLM. Выявляя закономерности в этом тексте, исследователи и разработчики учат LLM создавать текст самостоятельно, включая сообщения в блогах, стихи и компьютерные программы. Нейронные сети могут даже поддержать разговор.

ИИ-технологии способны помочь программистам, писателям и другим работникам генерировать идеи и действовать быстрее. Но доктор Бенджио и другие эксперты предупреждают также, что LLM могут научиться нежелательному и неожиданному поведению.

В частности, системы ИИ могут генерировать ложную, предвзятую и токсичную информацию. Такие модели GPT-4 неправильно воспринимают факты и выдумывают информацию – явление, называемое «галлюцинацией».

Компании работают над решением этих проблем. Но эксперты, и доктор Бенджио в их числе, обеспокоены тем, что по мере того, как исследователи сделают ИИ-системы более мощными, LLM создадут новые риски.

Краткосрочный риск: дезинформация

Поскольку LLM-системы предоставляют информацию в манере совершенной уверенности, при их использовании может оказаться непросто отделить правду от вымысла. Эксперты обеспокоены тем, что люди будут полагаться на эти системы для получения медицинских консультаций, эмоциональной поддержки и «сырой» информации, которую используют для принятия решений.

«Нет никакой гарантии, что эти системы будут корректны при выполнении любой задачи, которую вы им поставите», – сказал Суббарао

Камбхампати, профессор информатики в Университете штата Аризона.

Эксперты также обеспокоены тем, что люди будут злоупотреблять этими системами для распространения дезинформации. Поскольку LLM обучены разговаривать по-человечески, они могут быть удивительно убедительными.

«Теперь у нас есть системы, которые могут взаимодействовать с нами посредством естественного языка, и мы не можем отличить настоящее от подделки», – сказал доктор Бенджио.

Среднесрочный риск: потеря работы

Эксперты обеспокоены тем, что новый ИИ может уничтожить рабочие места. Сейчас такие технологии, как GPT-4, как правило, дополняют людей. Но OpenAI признает, что ИИ-технологии могли бы заменить некоторых работников, в том числе людей, модерирующих контент в Интернете.

Нейронные сети пока не могут дублировать работу юристов, бухгалтеров или врачей. Но они в состоянии заменить помощников юристов, личных помощников и переводчиков.

В документе, написанном исследователями OpenAI, говорится, что примерно для 80% рабочей силы в США программы LLM могут повлиять как минимум на 10% их рабочих задач, и что 19% работников могут увидеть, что затронуты по крайней мере 50% их задач.

«Есть признаки того, что рутинная работа исчезнет», – сказал Орен Этциони, исполнительный директор-основатель Института Аллена по искусственному интеллекту, исследовательской лаборатории в Сиэтле.

Долгосрочный риск: потеря контроля

Некоторые люди, подписавшие письмо, также считают, что искусственный интеллект может выйти из-под нашего контроля или уничтожить человечество. Но многие эксперты говорят, что это сильно преувеличено.

Письмо было написано группой из Института будущего жизни, организации, занимающейся изучением экзистенциальных рисков для человечества. Авторы предупреждают, что, поскольку на основе огромных объемов анализируемых данных системы искусственного интеллекта часто изучают неожиданное поведение, они могут создать серьезные и неожиданные проблемы.

Эксперты обеспокоены тем, что по мере того, как компании подключают LLM к другим интернет-сервисам, эти модели могут получить неожиданные возможности, поскольку смогут писать свой собственный компьютерный код. Они говорят, что разработчики создадут новые риски, если позволят мощным системам искусственного интеллекта запускать собственный код.

«Если вы посмотрите на простую экстраполяцию того, где мы находимся сейчас, на три года вперед, все будет чрезвычайно странно», – сказал Энтони Агирре, космолог-теоретик и физик из Калифорнийского университета в Санта-Крус и соучредитель организации «Институт будущего жизни». «Если вы возьмете менее вероятный сценарий – когда дела действительно развиваются, где нет реального управления, где эти системы оказываются более мощными, чем мы думали, – тогда все становится действительно безумным», – сказал он.

Доктор Эциони считает, что разговоры о экзистенциальном риске являются гипотетическими. Но он отметил, что другие риски – в первую очередь дезинформация – больше не являются спекуляциями.

«Теперь у нас есть серьезные проблемы», – сказал он. «Они непреднамеренные. Они требуют ответственной реакции. Они могут потребовать регулирования и законодательства».

К. Мец