МИР 24.02.2021 | Ева Вольфангель

Больше чем недоразумение

Искусственный интеллект подхватывает и запоминает расистские и сексистские предубеждения. Решить вопрос лишь технически невозможно.

В начале декабря известная исследовательница, технический руководитель команды по этическому использованию искусственного интеллекта в Google Тимнит Гебру сообщила в Twitter, что ее неожиданно уволили. За несколько часов эта новость всколыхнула все цифровое сообщество и вызвала волну солидарности с афроамериканской исследовательницей. Для многих ситуация ясна: Гебру стала неудобной из-за своей критики. Она утверждает, что не только внутри концерна меньшинства подвергаются расовой дискриминации, это происходит и посредством искусственного интеллекта (ИИ), создаваемого в нем или в других местах. Ее письмо с выражением протеста подписали более 2,5 тыс. сотрудников Google и свыше 4 тыс. представителей науки, поддерживающих ее взгляды.

Этот протест вынудил Джеффа Дина, главу подразделения по разработке ИИ, опубликовать электронное письмо, которое он отправил команде Гебру. В нем он заявил, что первопричиной конфликта является научная статья, написанная Гебру в соавторстве с другими учеными. Он не разрешил ее публикацию из-за того, что в ней недостаточно учтены результаты последних исследований. После этого Гебру стала угрожать заявлением об увольнении, которое было принято компанией в тот же день. В свою очередь Гебру утверждает, что не собиралась увольняться. Даже если произошло недоразумение, остается подозрение, что Google воспользовалась возможностью избавиться от неугодной сотрудницы.

И вот статья увидела свет: кроме Гебру, а также имен исследовательниц, не имеющих отношения к компании, например, ведущего автора Эмили Бендер из Вашингтонского университета, в ней не фигурируют другие сотрудницы Google. Критики едины во мнении, что статья сама по себе вряд ли может считаться основанием для увольнения. Если быть внимательнее, можно понять, почему

Google мешает позиция Гебру. Она считает, что концерн в абсурдной конкурентной гонке обучает ИИ на основе данных, присутствующих в обществе расовых и гендерных стереотипов.

Программа Google Translate постоянно переводит фразу «The doctor and the nurse» словами «врач и медсестра», хотя она с таким же успехом может обозначать «женщину-врача и санитара»

Авторы утверждают, что тенденция к созданию более крупных моделей и большего числа данных для обучения в компьютерной лингвистике приводит не только к использованию огромных ресурсов (энергия для работы серверных ферм). Таким образом, искусственный интеллект становится менее контролируемым, ущемляет права меньшинств, а разработчики не осознают этого.

Тут необходимо понять, как искусственный интеллект (ИИ) обучается языку. Системы машинного обучения получают от своих создателей два главных компонента: «данные на входе и желаемые данные на выходе». При машинном переводе ими могут быть, например, большие объемы текстов, доступных в различных языках. На их основе машины сами обучаются логическим межъязыковым связям. Интернет зарекомендовал себя при этом как особенно полезный источник для любой формы машинного обучения языкам, ведь в нем очень много языковых данных.

Но именно здесь и таится опасность, подчеркивают исследовательницы. По их мнению, большие массивы данных, основанные на текстах из Интернета, «будут иметь гегемонистическую точку зрения и закодированные предубеждения, которые могут ущемлять маргинальные группы населения».

И действительно в специальных разработках по машинному обучению в течение многих лет наблюдается проблема расистских и сексистских искажений. Об этом, в частности, свидетельствуют такие программы автоматического перевода, как Google Translate, закрепляющие ролевые стереотипы путем перевода профессий с языков без грамматической категории рода. Например, Google Translate переводит фразу «The doctor and the nurse» словами «врач и медсестра», хотя с таким же успехом можно обозначать «женщину-врача и санитара». Это происходит и в тех случаях, когда совершенно понятно, что речь идет о женщине-враче, потому что о ней уже упоминалось раньше.

Исследователи часто реагируют на дискриминацию по признакам расы или пола в системах искусственного интеллекта повышением количества данных для обучения

Исследователи часто реагируют на дискриминацию по признакам расы или пола в системах искусственного интеллекта повышением числа данных для обучения, а также увеличением соответствующих параметров в надежде на получение более репрезентативных результатов. Грубо говоря, параметры – это количество возможностей разграничения (дифференциации), которыми обладает такая сеть. Некоторые сравнивают их с числом синапсов в человеческом мозгу, а тем самым и выбором возможных комбинаций, что является весьма приблизительным отражением реальной картины.

Насколько мало пользы от огромного количества параметров для борьбы с человеческими предубеждениями при машинном обучении, показал пример языковой модели ИИ GPT-3 компании OpenAI. В ней – 175 млрд параметров, но и это не спасает ее от расистских и сексистских искажений, пишут сами разработчики и ученые OpenAI.

Гебру и ее коллеги допускают, что гигантские модели могут оказаться успешными в специфических случаях их применения. Однако в них преобладает нежелательный общественный эффект: такие модели удачно имитируют язык, не понимая его. Но их способность к созданию связных текстов отражает склонность человеческого мозга находить смысл в языке. Поэтому с этической точки зрения важно «исследовать в равной степени пользу и риск подражания человеку». Гебру и ее коллеги предлагают использовать модели поменьше, равно как и меньшие массивы данных, чтобы обеспечить время и мощности, необходимые для курирования их работы.

Решить вопрос лишь технически, в частности, уменьшением масштаба моделей и курирования массивов данных, невозможно

Подход Бендер, Гебру, а также их коллег, недостаточно радикален, считает Джоанна Брайсон, специалист по этике искусственного интеллекта Школы управления Герти в Берлине. «Многие полагают, что достаточно подобрать правильные данные для обучения, чтобы избавиться от предрассудков в машинном обучении», – говорит она. Но решить этот вопрос лишь технически, в частности, уменьшением масштаба моделей и

курирования массивов данных, невозможно. «Эти предубеждения – в нашей культуре, за ними стоим мы», – уверена Джоанна Брайсон. В 2017 году она вместе с другими исследовательницами, в том числе и человеческого мозга, показала, что похожие искажения существуют и в нашем образе мышления.

Неудивительно, что предвзятость, скрывающаяся в данных для обучения, заимствуется обучающимися машинами. В течение многих лет исследовательницы пытались устранить эту предвзятость путем изъятия фрагментов данных, но системы машинного обучения настолько хорошо находят закономерности в данных, что часто сами воссоздают расистский или сексистский контекст. Об этом свидетельствует известный пример программного обеспечения Compas, используемого судьями в США для принятия решений о досрочном освобождении заключенных. Система автоматически дискриминировала афроамериканцев, даже когда отсутствовала информация о цвете кожи. Благодаря журналистскому расследованию некоммерческой организации Pro Publica удалось выяснить, что программа вычисляла цвет кожи на базе третьей переменной, например, места жительства или имени.

Сорель Фридлер, специалист по информатике из США, которая занимается и вопросами этики, отмечает: «Невозможно удалить все атрибуты, которые могут привести к дискриминации, потому что для этого пришлось бы стереть почти все данные». К тому же люди многого и сами не осознают.

Предвзятость, искажения и несправедливость осознаются лишь при сравнении существующего статус-кво с идеальным обществом. Итак, не хватает формул справедливости. Вот только как они выглядят?

Предвзятость, искажения и несправедливость осознаются лишь при сравнении существующего статус-кво с идеальным обществом. И человек вначале должен втолковать это машинам в виде четких математических формул. Итак, не хватает формул справедливости. Вот только как они должны выглядеть? Возьмем относительно простой пример: у женщин и мужчин должны быть одинаковые шансы на рынке труда. Звучит вполне неоспоримо. Но как алгоритму распознать «равные шансы»? И можно ли считать неравной группу, в которой мужчины и женщины не составляют ровно половины?

Пойдем дальше: насколько иллюзорной является задача компенсации социальных предубеждений при помощи алгоритмов? И должно ли, например, программное обеспечение, осуществляющее автоматический отбор претендентов на собеседование, следить, чтобы мужчин и женщин было поровну? Или же ориентироваться на распределение всех заявок по признаку пола? А может, и полностью игнорировать его? И какие факторы еще должны быть учтены для обеспечения справедливого распределения? Квалификация? Цвет кожи? Возраст? И вот, казалось бы, простой пример растворяется в длинном списке вопросов.

У Брайсон есть иное предложение: «Нам нужно изменить свою культуру. Однако абсолютная корректность невозможна, ибо весь человеческий опыт основан на личном жизненном пространстве, и не существует возможности сбалансировать все отличия».

Неоднократно высказывались предположения, что Google заказывает лишь выгодные для рынка исследования. Нынешние события подтверждают эти подозрения.

Изменение нашей культуры – долгий, а может быть, даже нереальный путь. Но другой крайностью является то, что происходит в отделах по этике искусственного интеллекта таких крупных концернов, как Google. Неоднократно высказывались предположения, что Google заказывает лишь выгодные для рынка исследования. Нынешние события подтверждают эти подозрения.

Спустя всего несколько недель после запрета Джеффа Дина на публикацию критической статьи о больших языковых моделях с именем сотрудницы Google британская компания Google Deep Mind 11 января обнародовала рекорд: гигантскую языковую модель на 1,6 трлн параметров, которая прошла обучение на базе огромного массива данных объемом в 750 гигабайт, включающем тексты из Википедии, социального новостного сайта Reddit и других интернет-сайтов – «на два порядка больше, чем Википедия», заявили исследователи ИИ концерна Google. Для появления статьи, которая как раз подвергает критике это безумие, скорее всего, был неподходящий момент. В публикации не говорится ни о возможных негативных последствиях, ни о проблеме с искусственным интеллектом на базе общественных стереотипов, основанных на статьях в Интернете.

Мы будем далеки от справедливого будущего с искусственным интеллектом до тех пор, пока в исследованиях по разработке ИИ будем руководствоваться принципом «больше, дальше, быстрее».

Этика не должна сводиться лишь к формальным признаниям, которые ничего не стоят, как только перестают вписываться в маркетинговую стратегию.



Ева Вольфангель

Ева Вольфангель (Eva Wolfangel) – культуролог и журналист. В фокусе ее репортажей, которые печатаются в том числе в ZEIT, Spiegel и GEO, – технологии будущего, искусственный интеллект и виртуальная реальность.