АЛЬТРУИЗМ В ОБЛАСТИ ДАННЫХ

Благодаря стремительному прогрессу в области цифровых технологий, разработке различного рода датчиков и Интернета вещей в мире собирается все больше сведений из широкого диапазона областей, включающих жизнедеятельность человека, производственную сферу, природные явления, и формируется большой массив данных. В 2020 г. их объем достиг 44 зеттабайта. Бесспорно, такое обилие информации, и особенно при условии всеобщего доступа к ней, может стать основой для преодоления глобальных проблем и принятия обоснованных планов, решений и конкретных действий. Этой теме посвящена публикация руководящих принципов ЮНЕСКО «Открытые данные для искусственного интеллекта», лейтмотивом которых является утверждение о том, что государства – члены организации должны обмениваться контентом, обеспечивать его прозрачность и подотчетность, а также возможность использования для всех. Настоящие рекомендации направлены на то, чтобы способствовать общему пониманию важности открытых данных (ОД) для развития технологий искусственного интеллекта и проблем при работе с ними, а также для содействия цифровой трансформации и создания инклюзивных обществ знаний.

Экосистема управления открытыми данными

Конгломерат накопленных данных имеет огромный потенциал, но для его реализации они должны быть пригодными для обработки и внедрения результатов: обладать функцией быстрого поиска, открытого доступа, совместимости, модифицируемости и возможностью повторного использования. Поэтому к заинтересованным сторонам, особенно правительствам, а также научному сообществу и частному сектору, был обращен призыв предоставлять открытые данные. Первым, по утверждению авторов публикации, эту идею озвучил изобретатель всемирной паутины Т. Бернерс-Ли, обозначивший возможности использования ОД: «инновации, прозрачность, подотчетность, более эффективная управляемость и экономический рост». Актуальность этой тематики подчеркивалась в отчете ООН «Мир, который имеет значение» (2014 г.), в Рекомендациях ЮНЕСКО по открытой науке (2021 г.), в Законе ЕС об управлении данными, Альянсе больших данных для миграции (2021 г.), Директиве ЕС INSPIRE и в других документах.

Аналитики рекомендаций отмечают, что за последние годы инициативы в этой области значительно активизировались. Основан Институт открытых данных для формирования прозрачной и надежной экосистемы в 2012 г.; опубликовано Руководство по открытым правительственным данным и принята Хартия открытых данных G8 в 2013 г.; в 2015 г. подготовлена более инклюзивная версия Международной хартии открытых данных. Помимо этого в 2013 г. был создан Альянс по исследовательским данным, в 2014 г.– глобальная сеть «Открытые данные для развития», целью которой стало продвижение ОД для поддержки социальных изменений. В 2015 г. Рабочая группа по борьбе с коррупцией G20 опубликовала антикоррупционные принципы открытых данных, которые показывают их роль как инструмента обеспечения прозрачности, подотчетности и доступа к информации.

В 2018 г. Статистическая комиссия Департамента ООН по экономическим и социальным вопросам учредила Рабочую группу по открытым данным, занимающуюся выработкой принципов,

рекомендаций и поддержкой их внедрения в государствах – членах Организации Объединенных Наций. В 2020 г. Генеральный секретарь ООН представил Дорожную карту цифрового сотрудничества, а Координационный совет – Общесистемную дорожную карту по обновлению данных и статистики ООН, цель которой – способствовать улучшению мира посредством актуальной, достоверной информации, доступной для всех. В 2021 г. 193 государства – члена ЮНЕСКО приняли Рекомендацию по этике искусственного интеллекта – первый глобальный нормативный документ в этой сфере.

Исходя из утверждений авторов руководящих принципов, основным поставщиком больших объемов данных выступают государственные структуры, которые в прошлом ими практически не делились, к тому же ранее было невозможно получить их в функционально совместимом формате и использовать. В настоящее время картина кардинально изменилась: существует множество межправительственных инициатив и организаций, региональных структур и муниципалитетов в области ОД. Наиболее известно из них партнерство «Открытое правительство» – международная организация, учрежденная в 2011 г. с целью популяризации транспарентности государственного управления, подотчетного гражданскому контролю. Существуют также региональные платформы открытых данных, такие как openAFRICA, Азиатское партнерство, Латиноамериканская инициатива, инициативы G20 и Европейский портал открытых данных.

Огромное количество контента производится также в результате научной деятельности. И вот он, говорят эксперты, в первую очередь должен быть доступен для совместного использования, особенно если получен за счет государственного финансирования. Понимание важности открытой исследовательской информации было достигнуто давно – Комитет по данным Международного совета по науке был создан еще в 1966 г. Однако, несмотря на подписание Соглашения об обмене научными данными и работу различных платформ, получить доступ к ним трудно из-за большого объема, а также из-за некорректного управления ресурсами. Эту проблему пытаются решить с помощью Руководящих принципов FAIR, в которых делается упор на автоматический поиск данных, и Рекомендаций ЮНЕСКО об открытой науке.

Авторы публикации указывают на то, что производственная сфера и сектор услуг также накапливают масштабные сведения. Как правило, они связаны с профессиональной деятельностью, однако их значительная часть имеет отношение к устойчивому развитию. Раньше компании отказывались предоставлять свои данные, однако теперь появилась возможность делать это без риска для бизнеса. Несколько крупнейших мировых корпораций присоединились к инициативе UN Global Pulse под названием «Благотворительность в области данных» и делятся ими на благо общества, многие сотрудничают с Учебным и научно-исследовательским институтом Организации Объединенных Наций в рамках Программы оперативных спутниковых приложений ООН, чтобы обмениваться снимками, сделанными космическими аппаратами, в гуманитарных целях. У фирм есть множество стимулов для того, чтобы открыть свои данные, таких как возможность проведения анализа, доступ к новым идеям, улучшение репутации и связей с общественностью, увеличение дохода, соответствие нормативным требованиям, корпоративная благотворительность. В Законе ЕС «Об управлении данными» вводится механизм, называемый альтруизмом в области данных, который предоставляет возможности отдельным лицам и компаниям делиться имеющейся информацией для общественного блага и поощряет такие инициативы.

Совместимость и стандарты качества ОД

Наиболее точная формулировка этого феномена приведена Фондом открытых знаний: «Открытые данные и контент могут свободно использоваться, изменяться и распространяться кем угодно для любых целей». Поскольку основное внимание в настоящем руководстве уделяется открытым государственным данным, представляющим собой «философию – и все чаще набор политик, которые способствуют прозрачности, подотчетности и созданию стоимости путем обеспечения их доступности для всех», имеет смысл утверждать, что директивные органы не только собирают информацию и разрабатывают нормы для ее обмена, но также выступают в качестве ее поставщиков.

Важным дополнением к этому утверждению, отмечают аналитики руководящих принципов ЮНЕСКО, являются 6 основных принципов Международной хартии открытых данных по публикации ОД, разработанных в 2015 г. правительствами, гражданским обществом и экспертами по всему миру:

- *открытость по умолчанию;*
- *своевременность и полнота;*
- *доступность и удобство использования;*

- *сопоставимость и совместимость;*
- *улучшение управления и вовлечение в него граждан;*
- *инклюзивное развитие и инновации.*

Еще один важный набор характеристик, которым должны соответствовать ОД, называется FAIR. Эта аббревиатура означает находимые (Findable), доступные (Accessible), функционально совместимые (Interoperable) и повторно используемые (Reusable) данные, которые должны соответствовать следующим требованиям:

• для обеспечения возможности как ручного так и машинного поиска необходимо присвоить им глобальные уникальные и постоянные идентификаторы, разработать для них подробные метаданные и обеспечить их наличие в наиболее удобном для поиска ресурсе;

• для доступности крайне важно получать информацию по идентификатору через стандартизированный протокол связи – открытый, бесплатный, универсальный с процедурой аутентификации и авторизации;

• для функциональной совместимости следует представить данные в формате, пригодном для хранения, передачи и обработки;

• для повторного оборота данные должны быть хорошо описаны метаданными и соответствовать стандартам, а также иметь простую и доступную лицензию на их использование, информацию о происхождении, методе сбора и обслуживании.

Принципы FAIR были дополнены принципами CARE. Эта аббревиатура расшифровывается, как C – коллективная выгода, A – полномочия по контролю, R – ответственность и E – этика. Совокупность этих характеристик является краеугольным камнем всех действий по управлению данными, нацеленных на установление стандартов их обмена и этического использования.

Что касается технической стороны оборота ОД, то Т. Бернерс-Ли предложена пятизвездочная схема развертывания данных, для чего необходимо:

* сделать их доступными онлайн в любом формате;

** предоставлять их в структурированном виде (например, MS Excel вместо отсканированной таблицы);

*** выставлять в непатентованном открытом формате (например, CSV вместо MS Excel);

**** применять URI для обозначения объектов, чтобы другие пользователи могли указывать на них;

***** связать данные с другими для обеспечения контекста.

Чем больше звездочек, тем более открытыми являются данные.

С юридической стороны необходимо указать, как ОД могут повторно вовлекаться в оборот и требуется ли при этом указание авторства. Если данные не могут быть доступны по причине вопросов, связанных с неприкосновенностью частной жизни, например персональные или охраняемые, это следует четко оговаривать. Аккумулировать ОД необходимо в одном хранилище, например, собирая их из различных государственных ведомств, и заполнять пробелы в случае необходимости за счет сбора новых источников – точных, актуальных и полных. Данные о людях должны агрегироваться по доходу, полу, возрасту, расе, этнической принадлежности, миграционному статусу, географическому положению и т.д.

Настолько открыто, насколько возможно, настолько закрыто, насколько необходимо

Такого принципа рекомендуют придерживаться эксперты публикации в отношении открытых данных. В пользу первого утверждения приводится аргумент о том, что авторского права на фактологическую информацию в любом случае не существует, а доступ к ней должен иметь каждый. Кроме того, деятельность правительственных органов благодаря ОД становится прозрачной, что способствует укреплению доверия к государству. Возможность повторно использовать, перекомпоновывать и комбинировать их, а также потенциально получать из них новые идеи позволяет вовлекать граждан в процесс управления, а также создавать новые инновационные услуги и продукты, тем самым способствуя социальным, экономическим и экологическим реформам.

Противники открытого доступа утверждают, что сведения, доступные всем, могут нарушать конфиденциальность заинтересованного лица, которое вправе самостоятельно распоряжаться ими, а также авторские права или права на интеллектуальную собственность. К тому же во многих случаях сбор и трафик контента являются трудоемкими и дорогостоящими процессами, которые необходимо оплачивать, и если он будет распространяться без ограничений, от этих услуг будут отказываться. По мнению оппонетов, наборы данных могут содержать систематическую ошибку, возникающую на различных этапах, в том числе во время составления отчетов

человеком, отбора, маркировки и классификации, вплоть до интерпретации результатов моделирования. Существует также опасение, что открытая информация может быть использована не по назначению, со злым умыслом, особенно с учетом того, что многие новые технологии, включая ИИ, могут иметь двойное назначение.

Хотя ОД нужны для многих областей деятельности, эксперты подчеркивают их особую важность для генеративного искусственного интеллекта, представляющего собой успешные алгоритмы машинного обучения, в котором используются нейронные сети для понимания закономерностей и формирования новых данных, в том числе искусственных цифровых. Вместо простой обработки имеющихся сведений генеративный ИИ разрабатывает программы, которые могут производить новый контент, реалистичные изображения или видео, генерировать ответы на обычном разговорном языке, придумывать дизайн новых товаров и сочинять музыку. Существуют опасения по поводу возможного неправомерного использования генеративного ИИ, например для создания ложных новостей или дипфейков, вследствие чего необходимо разработать нормы ответственности за такие деяния и тщательно рассмотреть этические последствия его применения. Ведь качество продукта становится настолько высоким, что люди не могут различить, был ли он создан машиной или человеком, что вызывает опасения и споры. Поэтому авторы публикации обращают внимание на некоторые проблемы, существующие в отношении ИИ:

- *поток информации настолько велик, что огромную ее часть пока невозможно проанализировать;*
- *для работы отдельных инструментов ИИ не хватает данных или же они существуют в отрыве от остальных, являются устаревшими, ненадежными, неточными, имеют нерабочий формат или не размечены;*
- *в некоторых сферах имеются ограниченные, но критически важные сведения, однако пока нет алгоритмов ИИ, способных их адекватно проанализировать.*

Чем оперативнее обрабатываются данные, тем быстрее работают организации. Однако если скорость этого процесса снижается из-за проблем с доступностью контента, отсутствия необходимых инструментов и задержек, вызванных устаревшими версиями программ, компании могут не достичь желаемых результатов и утратить конкурентное преимущество. Известно, что необработанные данные бесполезны, пока не преобразованы из различных форматов в наиболее эффективный и пригодный для принятия наилучших решений.

Искусственный интеллект способен анализировать большие (или, если нет другого варианта, небольшие) объемы информации, выявлять ее ранее неизвестные или скрытые закономерности и предоставлять в режиме реального времени. Этот мощный ресурс должен использоваться, в частности, для реализации Целей устойчивого развития на период до 2030 г., а также для поиска решений глобальных проблем. Об этом говорится в Дорожной карте по цифровому сотрудничеству Генерального секретаря ООН. В ней содержится призыв укреплять потенциал для расширения цифрового взаимодействия и интеграции в это партнерство новых стран (цель 4), а также поддерживать глобальные проекты в области ИИ (цель 6).

Как избежать яда в колодце

Аналитики руководящих принципов обращают внимание на то, что искусственный интеллект – это технология двойного назначения, поэтому она связана не только с большими возможностями, но и с серьезными рисками. Наборы данных, которые загружаются в системы ИИ, могут привести к нежелательным результатам или реакциям. К тому же в отношении нейронных сетей могут совершаться и злоумышленные деяния, которые становятся результатом неправомерного использования ОД. Такие воздействия получили названия «яд в колодце» или «атаки уклонением» и включают в себя различные ухищрения. Одним из них может выступать, например, простая замена меток в наборах данных, тогда как другой, более изощренный способ связан с вредоносными образцами, которые тщательно продуманы, трудно распознаваемы и приводят к изменению исходного контента с целью получения неверного результата. Непреднамеренно повышает риск таких атак появление общедоступных наборов данных и рост их объема, потенциально доступного для опасных манипуляций, поэтому шансы выявить их снижаются.

Для использования ОД в системах ИИ они должны пройти соответствующую обработку, заявляют авторы публикации. В первую очередь форматирование – перечень процедур, которые помогут сделать данные более подходящими для машинного обучения. Их агрегация из разных источников или обновление вручную разными людьми может привести к нарушению последова-

тельности контента, поэтому он должен быть очищен и промаркирован, что зачастую требует много времени, а следовательно, и средств. Успех системы ИИ зависит от функциональности подготовленной информации, в том числе от ее согласованности и актуальности. Что касается необходимого объема обучающих данных, то, по мнению экспертов, его трудно предугадать заранее, но следует контролировать с помощью проверок производительности. И основное требование: ОД должны охватывать все сценарии, для которых была выстроена система искусственного интеллекта.

Первым шагом в открытии данных для ИИ аналитики руководящих принципов называют принятие решения о том, какие наборы следует сделать доступными. Критерии «за» следующие: существовали ли ранее запросы на них, была ли такая информация открыта и принесло ли это пользу, «против» – вопросы конфиденциальности, национальной безопасности, прав на интеллектуальную собственность и защиты персональных сведений.

Перед раскрытием наборов данных государство должно точно указать, на каких условиях их можно вводить в оборот. Из существующих вариантов, по мнению авторов публикации, оптимальным в этом случае является наиболее распространенный сегодня механизм лицензирования Creative Commons, предоставляющий возможность бесплатного, простого и стандартизованного способа получения разрешений на использование авторских прав на творческие и научные произведения, гарантирующих указание авторства и позволяющих другим копировать и распространять эти произведения. Для всех данных, регулирующихся лицензией Creative Commons, помимо отказа от всех прав, могут быть установлены следующие ограничения:

- атрибуция: указание источника;
- некоммерческий характер использования;
- запрет передачи данных для применения при их изменении;
- совместное использование обработанного контента на условиях той же лицензии.

Другой вариант, считают аналитики руководящих принципов, заключается в том, чтобы государства разрабатывали свои собственные лицензии, как это сделало, например, правительство Великобритании.

Наиболее распространенным способом открытия данных является их публикация в электронном формате для загрузки на веб-сайте, в архиве или репозитории, который должен, согласно Рекомендации ЮНЕСКО об открытой науке, «поддерживаться и обслуживаться академическим учреждением, научным обществом, государственным учреждением или другой хорошо зарекомендовавшей себя некоммерческой организацией, работающей на общее благо, что обеспечивает открытый доступ, неограниченное распространение, функциональную совместимость и долгосрочное цифровое хранение и архивирование».

Для этого существуют различные системы управления информационными потоками с открытым исходным кодом, позволяющие сделать портал ОД в виде одного окна. Наборы данных должны сопровождаться метаданными с универсальными стандартами Dublin Core, Data Catalog Vocabulary, DataCite100 и Schema.org, а форматы – поддерживать совместимость с интерфейсами прикладного программирования (API) и иметь возможность повторного использования.

ОД должны легко находиться по запросу. Для этого необходимо разработать коммуникативную стратегию, которая может включать анонсирование этого события в сообществах открытых данных и соответствующих каналах социальных сетей, например с помощью определенного хэштега. Помимо информирования следует стимулировать заинтересованных лиц к применению данных для решения конкретных задач и проблем, а также развивать международные партнерства для разработки систем ИИ. В то же время необходимо предотвращать злоупотребление ОД со стороны ненадежных и вредоносных приложений с использованием ИИ. Также рекомендуется вести публичный учет того, какие данные обрабатывались системами ИИ и каким образом.

Настоящие руководящие принципы – своеобразный призыв к действию в соответствии с Рекомендацией ЮНЕСКО об этике искусственного интеллекта. Если государства – члены организации будут следовать изложенной концепции, раскрывать информацию на постоянной основе и создавать для этого соответствующие возможности и культурную среду, то основные постулаты публикации могут стать реальностью.

Авторы призывают сделать данные доступными для поиска, совместимыми и многоразовыми, а также готовыми к взаимодействию с ИИ, словом, честными, чтобы их мог обрабатывать и анализировать кто угодно на благо общества. И

Ирина ЕМЕЛЬЯНОВИЧ

По материалам публикации ЮНЕСКО
«Открытые данные для искусственного интеллекта (ИИ)»
https://ipquorum.ru/upload/----4_korr_uid_64c8e0fa411cd-hpqhldfK.pdf.