УДК: 339.944.2

1.2. Экономические модели и недискриминируемый доступ к данным

Неволин И.В., к.э.н., в.н.с. ЦЭМИ РАН, Москва

Одним из ключевых элементов научной работы являются наблюдения и адекватные инструменты анализа исследуемых объектов. Общественные науки, и экономика в частности, оказались в невыгодном положении именно в части доступа к инструментам наблюдений и анализа. Бизнес в лице крупных ИТ-компаний накапливает детальную информацию социально-экономического характера в реальном времени. При этом научное сообщество лишено такой возможности: создание похожей инфраструктуры для решения бизнес-задач выходит за рамки научных целей и вряд ли найдёт соответствующую поддержку.

Говоря о развитии технологий искусственного интеллекта, Президент России озвучил ряд предложений по объединению усилий науки и бизнеса в этой сфере. Есть среди них и доступ бизнеса к данным государственных учреждений для поддержки машинного обучения. В статье приводятся аргументы в пользу того, что положения об обмене данными могут быть расширены для поддержки научных исследований об обществе и разработки компьютерных моделей для целей государственного управления.

Путешествие в мир искусственного интеллекта

Главной темой конференции 2023 года стали генеративные нейронные сети. Их широкое распространение в конце 2022 года в связи с появлением ChatGPT—2 и бурное развитие в части расширения функционала привлекли к себе большое внимание. Оно поддерживалось, во-первых, скоростью работы сетей, которые выдавали связные тексты и узнаваемые изобразительные образы. Во-вторых, скандалами вокруг результатов генеративных моделей. На это накладывались вопросы о перспективах творчества у человечества и поиски вариантов взаимодействия с искусственным интеллектом нового уровня. Высокая оценка технологии признаётся специалистами в области искусственного интеллекта: большой корпус данных позволил обучить нейронную сеть до уровня, который обеспечивает удовлетворительные результаты при широком наборе сценариев взаимодействия с человеком на понятном ему языке.

На заседании с участием Президента России В.В. Путина[1] продемонстрированы примеры достижений искусственного интеллекта в образовании, медицине, государственном управлении, представлена точка зрения разработчиков, в том числе, от той сферы, которая причастна к подготовке этих самых разработчиков. В ходе заседания прозвучали тезисы, которые следует интерпретировать как поручения Президента, пусть и не оформленные в виде соответствующих документов. Цель данной статьи – рассмотреть эти тезисы с точки зрения науки на предмет исполнения озвученных поручений для создания наилучших условий для развития технологий искусственного интеллекта в России. При этом, однако, анализ ограничивается одной сферой применения – государственным управлением, поскольку более общий взгляд потребовал бы формата, выходящего за рамки статьи.

Тезисы Президента

В выступлении и репликах Президента России хотелось бы отметить два блока тезисов. Первый касается инфраструктуры – научной, исследовательской, отраслевой – для развития технологий, разработки конкретных продуктов и приложений. Второй блок тезисов связан с последствиями и эффектами от использования искусственного интеллекта: вопросы ограничения разработок, недобросовестного использования технологии, этические вопросы. Поскольку Президент озвучил свою позицию на мероприятии, организованном корпорацией[2], уместно представить взгляд на инфраструктуру и последствия технологии со стороны научного сообщества. Точка зрения науки наряду с картиной, представленной на конференции бизнесом, позволит, как считается, сформировать более сбалансированный взгляд на развитие искусственного интеллекта. Да, на заседании выступал ректор университета Иннополис А. Гасников, но его тезисы затрагивали, главным образом, образование и развитие компетенций в регионах – тоже важного элемента инфраструктуры в части развития человеческого капитала. Однако он не обращался к информационной и вычислительной инфраструктуре, где имеется существенный дисбаланс, отмеченный Президентом. Поэтому остаётся место для уточнения позиции научного сообщества, и данная статья вносит свой посильный вклад в прояснение этой позиции. Далее следуют краткие обозначения тезисов и раскрывающие их цитаты.

[1] Стенограмма доступна по ссылке: http://kremlin.ru/events/president/news/72811 (дата обращения 15.12.2023)

[2] Имеется ввиду не конкретная организационно-правовая форма, а фактическое состояние Сбера, который имеет большой штат сотрудников, распределённую по территории деятельность, группу аффилированных лиц, управляет разнообразным по источникам дохода бизнесом.

Подготовка студентов и учащихся.

«Особые льготы для пользования вычислительной инфраструктурой должны получить аспиранты, студенты, школьники, которые уже занимаются научной и практической деятельностью в области искусственного интеллекта». «Нужно существенно расширить подготовку кадров, сильнейших учёных-разработчиков. Такую задачу необходимо ставить перед лидером первого рейтинга вузов по качеству подготовки специалистов в сфере искусственного интеллекта».

Доступ к суперкомпьютерам.

«*Прошу Правительство, Альянс, Российскую академию наук предложить механизм, который обеспечит доступ отечественных учёных к существующим и создаваемым в России суперкомпьютерам». «Обращаю внимание Правительства и компаний Альянса: действующие мощности российских суперкомпьютеров следует, конечно же, нарастить не менее чем на порядок».*

Важность корпуса данных

«Отмечу в этой связи совместный проект институтов Российской академии наук, "Яндекса", создавших колоссальную по объёму коллекцию текстов – Национальный корпус русского языка».

Обмен данными

«Считаю также необходимым, чтобы органы власти всех уровней, библиотеки и архивы помогали созданию наборов данных для машинного обучения, предоставляя их оперативно и безвозмездно. Я очень прошу коллег из Правительства продумать соответствующий механизм этой работы». «Добавлю также, что в наших планах создать хранилище кода отечественной платформы и сервисы, которые нужны для взаимодействия отечественных и зарубежных программистов».

Подготовка кадров затруднительна без практических занятий. И механизмы доступа к суперкомпьютерам, к корпусам данных направлены на то, чтобы сделать такие занятия более продуктивными. При этом отмечен положительный опыт взаимодействия науки и бизнеса в создании корпуса данных. Но обмен данными в озвученных тезисах выглядит однонаправленным: со стороны государства, учреждений культуры в сторону бизнеса. Для генеративных моделей это важно, что и было отмечено в выступлении, - обеспечить нейронную сеть качественным корпусом данных, и только специалисты в области культуры могут корректно подобрать и разметить документы для большинства сценариев повседневного использования искусственного интеллекта подобного уровня. Однако поток в обратном направлении может оказаться не менее ценным. Генеративные сети, как известно, обучаются не только на классических произведениях: корпус данных пополняется публицистическими материалами, энциклопедическими данными и т.п. – документами, которые, по мнению разработчика, являются наиболее характерными для пользователя. Взаимодействие с генеративной сетью, несомненно, оказывает определённое влияние на человека – не только с точки зрения информирования, но и в культурно-ценностном отношении. Здесь уместно привести результат конкретного эксперимента, выполненного М.А. Милковой. Перед ChatGPT поставлен вопрос: «Какой ученый писал о том, что элита состоит из тех, кто готов убивать и быть убитым? и что именно он писал?» Ответ сети: «Это высказывание принадлежит американскому философу и политологу Хорсту Найхау. Он написал: "Элита состоит из тех, кто смело решает убивать, чтобы сохранять свою власть, и умеренных, которые готовы умереть, чтобы сдерживать элиту"». Формулировка ответа предполагает отделение оппозиции от власти по признаку готовности к жертвам. Для власти жертва – естественный спутник деятельности. Для оппозиции – вынужденное и побочное явление деятельности. И это представление транслируется неискушённому пользователю со ссылкой на конкретного учёного, в трудах которого не удалось найти такой позиции! В связи с этим примером (и многими другими, включающими обращения адвокатов за помощью к генеративным моделям для подготовки материалов по конкретным судебным делам) возникают вопросы об оценке выхода нейронных сетей, об анализе пользовательских реакций. Здесь открывается огромное поле для научных исследований и академической экспертизы, которые, однако, не реализуемы без доступа к данным. Вряд ли этот доступ должен быть свободным, но сама возможность для научного сообщества получить такие данные, обсуждение механизма такого доступа оказали бы большую услугу государству.

Этика

Говоря об этике применения искусственного интеллекта, Президент обратил внимание на ущерб от внедрения искусственного интеллекта. В том числе, указал на специфичный подбор документов для обучения нейронных сетей. Фактически, озвучена проблема недобросовестной конкуренции: крупные игроки могут навязывать свою точку зрения. Это важная проблема, поскольку обыватель склонен считать, что машина действует (и выдаёт материалы) непредвзято, и это отличает её от человека-

редактора. В действительности человек влияет на самое главное – на подбор документов и их интерпретацию в ходе последующей разметки. Выше, при описании преимуществ для государства от доступа научного сообщества к данным корпораций, об этом частично сказано. Монополия над данными – результаты взаимодействия с искусственным интеллектом, протоколы поведения клиентов, в том числе, экономического – со стороны крупных частных корпораций имеет схожие с недобросовестной конкуренцией признаки. Большое сообщество учёных общественного профиля развивает теорию, проверяет её расчётами на доступных (часто агрегированных и переработанных) данных, в то время как несколько лиц контролируют детальную информацию об обществе, монополизируя право на её анализ и интерпретацию. Таким образом, обмен данными между корпорациями, наукой и государством имеет и этическую сторону – добросовестное ведение бизнеса с поддержкой исследований нашего общества, в том числе, для целей государственного управления.

Выход за рамки генеративных сетей. Государственное управление

Говоря о распространении нейронных сетей, Президент выступил с предложением

«подумать о разработке на основе генеративного искусственного интеллекта больших отраслевых моделей, предложить механизмы их практического внедрения, чтобы существенно повысить производительность труда, а значит, и заработные платы в ключевых областях отечественной экономики». Разработка моделей для отраслей, для государственного управления требуют анализа и контроля программного выхода, и тут предложения о расширении доступа к данным корпораций более чем уместны.

Если посмотреть на ситуацию шире, технологии искусственного интеллекта не ограничиваются генеративными моделями, которые действительно громко заявили о себе в 2023 году. На одном из последующих мероприятий – Встрече Президента России с молодыми учёными – об этом прямо заявил М. Криницкий[3]. Соответственно, решение проблемы подготовки кадров, развития науки и принятие методов искусственного интеллекта научным сообществом затрагивает и обратный поток данных: из отраслей в науку. Какие данные собирают компании? Как устроены измерения/ наблюдения? Насколько принятые научным сообществом методические вопросы анализа данных того или иного типа соответствуют реально проводимым отраслями измерениям? На эти и многие другие вопросы необходимо дать чёткий ответ, не позволяя промышленности (практике) и науке (теории) разойтись в используемых методах на недосягаемое расстояние. В противном случае попытка найти нужного специалиста (группу, институт) для решения конкретной задачи производства (фактически, задачи повышения эффективности экономических процессов) будет обречена на неудачу. И это – обмен данными в обе стороны – следует иметь всегда, когда поднимается вопрос о взаимодействии науки и бизнеса.

Как генеративные модели вписаны в общий контекст технологий искусственного интеллекта, так и данные, связанные с такими моделями, вписаны в более широкий контекст того, что мы знаем об обществе, что можем измерять. Какие данные доступны учёным общественных наук? Статистические агрегаты, индексы для описания картины в целом и выборочные обследования для изучения конкретных взаимосвязей на индивидуальном уровне. При этом агрегаты (такие как макроэкономические показатели) представляют собой переработанную информацию, которая публикуется с задержками. Ещё и с пересмотром методики расчётов, что затрудняет межвременные сравнения. Индивидуальные обследования всегда поднимают вопросы репрезентативности выборки, которые выходят далеко за пределы определённого уровня разнообразия по полу и возрасту респондентов. Те же самые выборочные обследования используются и на уровне Росстата, например, при исследовании доходов населения. Операции с экономическими агрегированными показателями и выборочное обследование в науке и госуправлении выглядят неадекватными на фоне имеющейся детальной и сплошной (в смысле охвата населения) информации у корпораций – о потреблении (покупках), доходах (заработных платах), миграции (геолокации в динамике) и т.п.

Социально-экономические модели для целей государственного управления опираются, в том числе, на агент-ориентированный подход. В таких моделях возможно описание отдельного гражданина, отдельного предприятия, отдельной единицы транспорта, товара и т.д. Суперкомпьютерные технологии поддерживают численное моделирование задач большой размерности: потребления социальных услуг города его жителями, развития человеческого капитала страны (исследование образовательных и карьерных траекторий в науке), распространённости заболеваний и т.д. Модели эпидемий в последнее время получили широкую известность в контексте динамики COVID-19. Однако возможны иные конструкции. Так, питание, очевидно, является важным компонентом здоровья, но описание конкретных функциональных связей между потреблением продуктов и развитием заболеваний по-прежнему актуально. В медицинских исследованиях проводятся контролируемые эксперименты по связи диеты со здоровьем. Но для подготовки экспериментов (тоже выборочных обследований) полезным оказывается статистический анализ данных, которые имеются, в том числе и у корпораций. В конкретном исследовании распространённости анемии операторы фискальных данных отказали в доступе к обезличенным данным о покупках. Выручили выборочные обследования RLMS (Russian Longitude

[3] Стенограмма доступна по ссылке http://kremlin.ru/events/president/news/72869 (дата обращения 15.12.2023)

Monitoring Survey) (НИУ ВШЭ, 2022) и собственный мониторинг цен в торговых точках в региональном разрезе. Эти данные, менее богатые и более дорогие для сбора, чем в случае корпораций или операторов фискальных данных, всё же позволили сделать количественные оценки связей между потреблением конкретных продуктов и развитием анемии (Дукхи и др., 2022), а также построить агентную модель для анализа рисков развития заболевания в результате изменения доступности сбалансированного рациона в регионах нашей страны (Машкова и др., 2021).

Прошедший Всероссийский конкурс моделей для государственного управления выявил большое количество отечественных разработок (Неволин, 2023). На макроскопическом уровне все они опираются на статистическую информацию (устаревающую и агрегированную, как отмечалось выше). На уровне предприятий успешные модели построены по результатам взаимодействия с фирмами, заинтересованными в решении стоящих перед ними задач. Ожидается, что и государство получит ощутимую пользу от решения задач на муниципальном, региональном, национальном уровнях при использовании детальной информации, собираемой, в том числе, корпорациями.

Конкуренция с провайдерами услуг и разработчиками моделей

Доступ научного сообщества к собираемым корпорациями данным может натолкнуться на возражение следующего рода. Компании вкладывали собственные ресурсы для накопления этих данных, имеют право на монопольный доступ к ним, а расширение доступа создаст неравные условия по разработке сервисов на их основе, по выводу информационных продуктов на рынок, поскольку конкурент не понёс соответствующих издержек и, таким образом, не отягощён дополнительными обязательствами. Возражение выглядит закономерным, если не вдаваться в детали данных и перечня лиц, которым открывается доступ.

Во-первых, рассмотрим суть накопленных данных. Они составляют информацию об экономическом и социальном поведении большой группы лиц – едва ли не всех граждан страны. Монопольный доступ к таким данным со стороны частных корпораций вызывает вопросы. Безусловно, существуют акты о доступе специальных служб и ведомств к таким данным. Но, как сказано выше, отечественная наука общественного профиля рискует застрять в неадекватной картине нашего общества. Для многих отраслей экономики доступны данные о выпуске продуктов – в разрезе моделей подвижного состава, например, или других характеристик. О потреблении продуктов и услуг в цифровой экономике таких детальных данных нет. Поэтому предлагается, реализуя поручения Президента, также продумать механизм получения этих данных.

Во-вторых, обратимся к конкуренции. Речь идёт именно о возможности научных исследований. Корпорации быстрее в выпуске продуктов, имеют профильные отделы, которые отвечают за улучшение пользовательского опыта, имеют несравнимый с научными организациями ресурс для продвижения своих продуктов. Это означает, что даже если какая-то из научных организаций задумает предложить свой коммерческий продукт на основе данных, он имеет низкие шансы на успех. Главный продукт науки – кадры высшей квалификации, научные публикации, научные мероприятия. Да, научные организации оказывают услуги (по проведению научных исследований) в рамках хозяйственных договоров. Но в целом, продукты и услуги показывают нацеленность на разные аудитории потребителей научного знания и информационных сервисов. В некотором узком сегменте деятельности корпораций – аналитических исследованиях – действительно может возникнуть конкуренция, но те, кто при этом обращается к научным организациям, идёт туда за брендом – за правом аргументировать свою позицию фразой «подтверждено научными исследованиями».

В-третьих, доступ к данным необязательно может быть свободным от обязательств. Известны различные типы лицензий на программные продукты – например, для некоммерческого использования частным лицом, академическая лицензия, лицензия для бизнеса и т.п. Каждая из них учитывает возможности и цели доступа к программным продуктам, накладывает соответствующие ограничения на функционал. Известны консорциумы организаций, которые, в том числе, разделяют общие цели и координируют свою деятельность. Полезным для научных исследований оказался бы такой механизм, который чётко устанавливал критерии доступа к социально-экономическим данным, хранящимся у корпораций.

Безопасность

Следующее возражение, которое можно ожидать в качестве препятствия к доступу, связано с вопросами безопасности – коммерческой инфраструктуры, коммерческого продукта, граждан и государства на всех уровнях управления. Если данные открываются широкому кругу организаций, повышается риск их утечки и недобросовестного использования информации о покупках, передвижении, социальных взаимодействиях граждан. Ответ на это возражение также можно прорабатывать, имея в виду несколько пунктов.

Во-первых, практика обезличивания данных не является новой и уникальной. Здесь можно сослаться на результаты выборочных обследований. Каждая анкета фактически представляет конкретного индивида/ конкретное домохозяйство. Богатая и доступная коллекция таких данных собрана при повторяющемся обследовании RLMS. Наборы данных включают доходы и расходы, образование и занятость, медицинскую часть. И до сих пор неизвестны случаи деанонимизации респондентов и злоупотребления информацией.

Во-вторых, механизм, который позволил бы научным организациям получить доступ к данным, и не должен предполагать свободную передачу информации. Уже упомянутые данные RLMS доступны исследователям только после регистрации на информационном ресурсе проекта. Это самый мягкий барьер. Более сложные барьеры могут устанавливать строгие, но обоснованные требования к организациям, предполагающие, например, технологический аудит информационной инфраструктуры и политики обращения с конфиденциальной информацией у претендента. Здесь уместно напомнить об использовании медицинских данных для обучения искусственного интеллекта и об использовании соответствующих сервисов медицинскими учреждениями, о чём было сказано на упомянутой конференции «Путешествие в мир искусственного интеллекта». Этот прецедент работы с чувствительными данными признаётся успешным, и он вселяет уверенность в то, что и для использования научным сообществом социально-экономических данных, собираемых корпорациями, также может быть найден подходящий механизм.

Заключение

Складывающаяся ситуация в информационной сфере, в цифровой экономике усиливает разрыв между наукой и бизнесом. Первая развивает модели, в том числе, компьютерные, на основе агрегированной и устаревшей информации. В таких условиях возможности по исследованию социально-экономических эффектов по своему охвату и глубине анализа не удовлетворяют потребностям, которые существуют в государственном управлении. Бизнес, в свою очередь, справедливо рассматривает данные как ресурс. И они становятся таковым только в случае придания им свойства редкости – путём ограничения доступа.

Президент России однозначно высказался за совместную работу науки и бизнеса в рамках развития систем искусственного интеллекта, обозначил направления, которые должны способствовать такой работе. К ним можно добавить предоставление доступа для научного сообщества к данным, собираемым крупными ИТ-компаниями о транзакциях в экономике. Поручения Президента направлены на разработку механизмов – готового решения нет, но старт уже дан. В этом отношении вопрос о доступе к упомянутым данным не хуже уже поставленных: он тоже требует проработки, и также важен для развития страны.

Литература

1. Дукхи Н., Бурилина М.А., Машкова А.Л., Неволин И.В., Сьюпол Р. (2022). Эконометрическое исследование пищевых факторов профилактики анемии // Вестник ЦЭМИ.– № 1. – DOI 10.33276/S265838870018349-6
2. Машкова А.Л., Дрипта Р.Ч., Ришемжит К., Неволин И.В. (2021). Агент-ориентированная модель взаимосвязи доступности продуктов питания и динамики распространения анемии // Искусственные общества. – Т. 16, № 1. – DOI 10.18254/S207751800013573-9
3. Неволин И.В. (2023). Искусственные общества: технологии построения и сферы применения // Экономическая наука современной России (в печати)
4. НИУ ВШЭ (2023). Национальный исследовательский университет "Высшая школа экономики" - Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ // НИУ ВШЭ: сайт. - [Б. м.], 1993-2022. -URL: http://www.hse.ru/rlms (дата обращения: 12.12.2023).

References in Cyrillics

1. Dukhi N., Burilina M.A., Mashkova A.L., Nevolin I.V., S'jupol R. (2022). Jekonometricheskoe issledovanie pishhevyh faktorov profilaktiki anemii // Vestnik CJeMI.– № 1. – DOI 10.33276/S265838870018349-6
2. Mashkova A.L., Dripta R.Ch., Rishemzhit K., Nevolin I.V. (2021). Agent-orientirovannaja model' vzaimosvjazi dostupnosti produktov pitanija i dinamiki rasprostranenija anemii // Iskusstvennye obshhestva. – T. 16, № 1. – DOI 10.18254/S207751800013573-9
3. Nevolin I.V. (2023). Iskusstvennye obshhestva: tehnologii postroenija i sfery primenenija // Jekonomicheskaja nauka sovremennoj Rossii (v pechati)
4. NIU VShJe (2023). Nacional'nyj issledovatel'skij universitet "Vysshaja shkola jekonomiki" - Rossijskij monitoring jekonomicheskogo polozhenija i zdorov'ja naselenija NIU VShJe // NIU VShJe: sajt. - [B. m.], 1993-2022. -URL: http://www.hse.ru/rlms (data obrashhenija: 12.12.2023).

Неволин Иван Викторович, к.э.н., в.н.с. ЦЭМИ РАН (i.nevolin@cemi.rssi.ru)
ORCID: 0000-0002-8462-9011

Ключевые слова

искусственный интеллект, большие данные, экономические исследования, экономико-математические модели, государственное управление

Ivan Nevolin. Economic models and non-discriminated access to data

Keywords

artificial intelligence, big data, economic research, mathematic models in economics, public administration

DOI: 10.34706/DE-2023-05-02

JEL classification C80 – общие положения о методологии сбора и оценки данных, компьютерных программах, A21 – связь экономики с другими дисциплинами.

Abstract

Scientific work relies strongly on observations and appropriate tools to analyze the objects under study. Social sciences – and economics in particular – are placed at a disadvantage position just because of access to observations and analysis tools. Business embodied the large IT companies accumulates detailed socio-economic information in real time. Meanwhile, the scientific community lacks such an opportunity: the creation of a similar infrastructure for solving business problems goes beyond scientific goals and is unlikely to find appropriate support.

While speaking about the development of artificial intelligence technologies, the President of Russia announced a number of proposals to combine the efforts of science and business in this area. Among them is access to government data for business enterprises in order to support machine learning. The article provides arguments in favor of the fact that the points on data exchange could be expanded to support scientific research about society and to sponsor the development of computer models for public administration purposes.