DOI: 10.17323/2587-814X.2022.2.62.73

Доверенный искусственный интеллект как способ цифровой защиты

С.М. Авдошин
E-mail: savdoshin@hse.ru

Е.Ю. Песоцкая
E-mail: epesotskaya@hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: Россия, 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация

Статья посвящена аспектам, связанным с феноменом цифровой защиты и массового внедрения нейронных сетей в современном обществе. С одной стороны, проникновение цифровых технологий расширит возможности бизнеса и социальных институтов, а также общества, которым поручено выполнять многочисленные задачи с большей эффективностью. Однако есть и серьезные угрозы, связанные с использованием искусственного интеллекта в части работы с данными пользователей. Авторы исследуют вызовы и формулируют новую концепцию надежного искусственного интеллекта для противодействия некорректному цифровому проникновению. Целью статьи является изучение потенциальных опасностей злоупотребления нейронными сетями со стороны владельцев технологий и сопротивления им с опорой на доверенный искусственный интеллект. Изучая различные аспекты цифровой трансформации и использования технологий искусственного интеллекта, авторы формализуют опасности и предлагают подход к использованию технологий цифровой защиты, которым можно доверять.

Ключевые слова: нейронные сети, цифровая защита, доверие, искусственный интеллект, общество

Цитирование: Авдошин С.М., Песоцкая Е.Ю. Доверенный искусственный интеллект как способ цифровой защиты // Бизнес-информатика. 2022. Т. 16. № 2. С. 62–73. DOI: 10.17323/2587-814X.2022.2. 62.73

Введение

Искусственный интеллект (ИИ) занимает все большее место в современном обществе, становясь основой для инноваций и прогресса [1–5]. Не являются исключением сферы, активно внедряющих цифровые технологии с целью модернизации процессов и ускорения экономического развития.

Концепция искусственного интеллекта (ИИ) была введена Джоном Маккарти в 1955 году [6, 7]. Понятие ИИ подразумевает способность системы интерпретировать данные, извлекать из них уро-

ки и использовать полученные сведения для достижения конкретных целей и задач посредством гибкой адаптации [6].

Стоит отметить, что обсуждение определения ИИ к настоящему времени не привело к результату, удовлетворяющему всех заинтересованных сторон в технологиях ИИ. С 2016 года индустрия искусственного интеллекта развивается благодаря поддержке облачных вычислений и больших данных. Сегодня искусственные нейронные сети, развернутые с использованием облачных технологий, а также глубокое машинное обучение формируют основу большинства приложений, которые нам известны под названием ИИ.

Использование таких сетей наряду с компьютерными мощностями даёт возможность решать различные проблемы, поскольку искусственные нейронные сети состоят из абстрактной модели взаимосвязанных нейронов, специальное расположение и связь которых могут быть использованы для решения прикладных задач в различных областях, таких как статистика, технологии или экономика [8].

Аналитические и когнитивные инструменты ИИ позволяют владельцем технологий анализировать колоссальные по объему данные, а значит своевременно выявлять и эффективно реагировать на изменения в ситуации в бизнесе и обществе. Именно нейронные сети могут стать важным шагом на пути к реализации принципов открытого общества, повысить уровень прозрачности, оптимизировать внутренние процессы межведомственного взаимодействия и стимулировать инновационную деятельность, что в итоге приведет к высокому уровню доверия [9].

На повестке дня для многих компаний и организаций стоит вопрос создания и последующего внедрения централизованной цифровой экосистемы, направленной на улучшение взаимосвязи и взаимодействия всех заинтересованных лиц (государства, бизнеса, объединений и частных лиц) [9–11]. Главная роль в этой трансформации будет возложена именно на ИИ.

Однако ситуация неоднозначна и ставит под вопрос своевременность широкого использования ИИ [12]. Помимо уже всесторонне определенных рисков нарушения приватности, хакерских атак, технологической сингулярности и некорректного цифрового проникновения, имеют место конкретные предпосылки другой серьезной опасности [5, 13–15]. Речь идет об аккумулировании и последовательном использовании возможностей цифровых технологий в интересах узкой группы людей. Однозначно можно утверждать, что набор современных алгоритмов открывает практически безграничные возможности. Подробный сценарий отражает концепцию появления в современном мире цифровой диктатуры (digital dictatorship) и необходимости цифровой защиты для регулирования недобросовестного использования технологий.

Выходом может стать повсеместное установление принципов доверенного ИИ (Trusted Artificial Intelligence) [6, 12]. В рамках этой концепции предполагается найти золотую середину между прогрессивностью стратегий применения ИИ и защитой этических и моральных аспектов человеческой жизнедеятельности. При имеющейся незначительной законодательной базе и в отсутствие международного консенсуса применение нейронных сетей на данный момент – это не только вызов традиционной модели работы властей, но и угроза основным человеческим ценностям [6, 7, 12]. Более того, имеется огромный разрыв между основами международного права и реальностью, где эти принципы постоянно нарушаются. Крайне важно выстроить продуманную нормативно-правовую систему, отвечающую запросам научно-технического прогресса и регулирующую все многообразие алгоритмов нейросетей. Важно также разработать конкретные технологии, которые позволят противодействовать возможным угрозам.

Таким образом, целью данной статьи является изучение рисков злоупотребления возможностями ИИ со стороны владельцев технологий и способов сопротивления им с опорой на концепцию доверия. Чтобы избежать проблем, которые могут нанести вред человеку путем искажения, хищения или утечки данных, необходимо убедиться, что результатам работы ИИ можно доверять. В данной работе исследуются потенциальные проблемы, связанные с вопросами доверия, конфиденциальности и надежности, рассматривается понятие «доверенного искусственного интеллекта» а также феномен цифровой защиты. В разделе 1 подробно разбирается природа нарушений прав граждан с точки зрения четырех различных сфер жизни: политической, социальной, культурной и экономической. Мы указываем на необратимость процесса внедрения нейронных сетей в контексте цифровой трансформации и описываем идею доверенного ИИ.

Раздел 2 посвящен обсуждению возможных стратегий реагирования и предложения по установлению цифровой защиты; завершается наше исследование кратким изложением ключевых выводов.

1. Перспективы и методы искусственного интеллекта

1.1. Новые возможности искусственного интеллекта

Научно-технический прогресс не остановить. Под давлением стремительного развития IT-сектора мировой экономики государства оказываются вынуждены отвечать меняющимся запросам бизнеса, в противном случае рискуя столкнуться с настоящей цифровой пропастью в управлении [9]. Искусственный интеллект плавно проник в нашу жизнь, умные системы применяются во множестве сфер, от клиентской аналитики и поисковых систем до голосовых помощников и медицинских исследований. В медицинской области разрабатываются системы, распознающие патологии по видеозаписям эндоскопических исследований; в транспортной – автопилоты и системы управления транспортным потоком; в финансовой – системы, опознающие клиентов или выявляющие подозрительные операции, которые могут указывать на уклонение от налогов или отмывание денег. Можно с уверенностью сказать, что дальнейшее мировое экономическое развитие и прогресс напрямую зависят от того, насколько эффективно различные индустрии научатся пользоваться искусственным интеллектом. Однако с развитием технологии более остро встаёт проблема доверенного искусственного интеллекта, поскольку любые неполадки могут иметь тяжелые последствия. Пользователи хотят быть уверены, что модель обладает высокой точностью, а ее результаты справедливы, легко интерпретируемы. Например, неверно откалиброванные датчики автомобиля, оборудованного автопилотом, могут стать причиной аварии. Ошибки в контроле над инфраструктурой, на которую полагается ИИ, могут привести к утечке персональных данных пациентов или краже личности. Низкий уровень программного обеспечения (ПО) ИИ может привести к постановке неверного диагноза или задержке в транспортировке, разрушающей дальнейшую цепочку поставок.

С одной стороны, вычислительные и аналитические мощности нейронных сетей, превосходящие возможности человеческого мозга, открывают новые горизонты для бизнеса и социальных институтов (например, в предиктивной аналитике). Помимо этого, ИИ не имеет ограниченного запаса выносливости и доступен для использования в любой момент. Благодаря этим характеристикам возможны колоссальные преобразования во многих секторах экономики: системах дорожного движения, здравоохранении, поддержании общественного порядка и персонализации государственных услуг, в том числе образования [15]. Внушительные успехи уже достигнуты в предоставлении услуг гражданам и юридическим лицам. При этом, конечно, важно отметить ограниченность выполняемых ИИ на данный момент задач. Среди основных категорий – проработка запросов (о социальных выплатах, миграции, по вопросам граждан и т. д.), заполнение и поиск документов, перевод текстов [4].

Значительные успехи нейронным сетям пророчат и в области управления экономикой. Планируется качественное повышение эффективности в распределении ресурсов и логистики через переустройство и оптимизацию каналов поставок, систем складов и утилизации [15]. Также нельзя не упомянуть многообещающую концепцию «Умный город» (smart city), опирающуюся именно на нейронные сети в вопросах контроля камер видеонаблюдения, электрических сетей, водоснабжения, транспортных систем и т. д. [7, 16].

Нужно отметить, что значительные изменения произошли под влиянием пандемии COVID-19 – в частности в области преподавания и обучения. Академические учреждения переходят на цифровые технологии, чтобы предоставить своим студентам больше ресурсов. Благодаря технологиям у студентов теперь есть больше возможностей учиться и улучшать навыки в своем собственном темпе и по индивидуальной траектории, проходить этапы контроля при помощи онлайн-тестов. Все больше популярности набирают сервисы онлайн-прокторинга, при которых лицо испытуемого идентифицируется и анализируется, чтобы предсказать его эмоции. Кроме того, обнаруживаются такие аспекты, как телефон, книга или присутствие другого человека. Эта комбинация моделей создает интеллектуальную систему логического вывода, основанную на правилах, которая способна определить, имело ли место какое-либо злоупотребление во время прохождения экзамена или теста.

Здесь может вставать вопрос о корректности работы системы и адекватности оценивания поведе-

ния. Любые сбои и злоупотребления чреваты негативными последствиями в части академической честности, дискредитации идеи как офлайн, так и онлайн-обучения. Среди ключевых рисов стоит отметить нарушение конфиденциальности, скомпрометированная доступность или скомпрометированные учетные записи, утечка персональных данных или искажение результатов.

Для успешной разработки моделей машинного обучения необходимы большие данные. Количество, качественно и доступность больших данных влияют на эффективность и точность обучаемых моделей, поэтому многие компании заинтересованы в организации постоянного сбора данных о своих потребителях. Многие системы собирают информацию, не подлежащую разглашению: видео и фотографии с видеокамер, записи речи, финансовые транзакции. Неразумное использование этой информации, ошибки в модели или хищение данных могут стать причиной угрозы безопасности отдельных личностей или даже предприятий и государственных организаций.

1.2. Потенциальные проблемы и угрозы цифрового проникновения

Утечка данных для искусственного интеллекта особенно опасна из-за того, что большие данные обычно несут в себе очень много конфиденциальной информации, из которой можно получить много сведений об объекте, на который осуществлена атака. При этом утечка данных может произойти на любом этапе разработки: обучении или использовании готовой модели.

Нарушение конфиденциальности является другой важной деталью, поскольку именно личная информация выступает катализатором цифровой трансформации, становясь основой для моделей обучения. Процесс разработки многих нейронных сетей практически неотделим от опоры на собираемые данные включая речь, интернет-активность, изображения, финансовые потоки, медицинские показатели и т. д. Таким образом, вопрос доступа к большим данным оказывается одной из самых существенных проблем, связанных с интеграцией нейронных сетей в общество [13]. Любой третьей стороне будут очевидны потенциальные риски использования собираемой владельцами технологий информации касательно информации пользователей, их права на частную жизнь и защиты их персональных данных [17, 18]. Этот аспект широко обсуждается обществом, так как за ним стоит прямая угроза здоровью и жизни человека. Актуальность проблемы многократно озвучивалась видными научными и государственными деятелями, а соответствующие отчеты и исследования не раз были представлены в рамках международной дискуссии (IEEE, специализированных комитетов ЕС, ОЭСР и т. д.) [6].

Публично гарантируя прозрачность, полный аудит и объективность благодаря внедрению нейросетей, владельцы технологий в эпоху цифровизации способны осуществлять любые манипуляции. Данная возможность вытекает напрямую из непонимания обществом и пользователями цифровых технологий структуры и принципов работы алгоритмов. Сложные модели искусственного интеллекта выполняют колоссальные вычисления, которые не поддаются всестороннему анализу даже разработчиков [5, 7]. Таким образом, для большинства пользователей процесс работы нейронных сетей будет непрозрачным [15]. Ученые называют это проблемой «черного ящика». Пользуясь этим феноменом, недобросовестные разработчики могут использовать нейронные сети по своему усмотрению.

Еще одна проблема, связанная с данными для обучения, это их низкая доступность. Зачастую небольшие объемы данных принадлежат разным лицам, не имеющим оснований доверять друг другу или разработчику, и составить один датасет – набор данных – достаточного размера оказывается невозможным. Если же датасет необходимого объема существует, он все равно может быть недоступен, если данные в нем конфиденциальны. Даже если подходящие данные удается найти, необходимо удостовериться, что они отражают реальное положение вещей. В частности, в них не должно содержаться скрытых предвзятостей, как это, например, произошло с некоторыми системами распознавания лиц: из-за несбалансированности данных они, например, значительно лучше справлялись с распознаванием светлокожих мужчин, чем темнокожих женщин.

Данные, с которыми работает готовая система ИИ, также могут быть низкого качества: они могут исходить из ненадежных источников или содержать информацию с высокой степенью неопределенности. Кроме того, базы данных, с которыми взаимодействует система, могут оказаться под угрозой, если сама система взломана. Например, система проверки биометрических данных может быть

подвержена нескольким типам атак с целью заставить ее принять злоумышленника за владельца. В обрабатываемые данные добавляется шум так, чтобы уже обученная модель определила объект на фотографии в неправильный класс. Такие атаки могут применяться в компьютерном зрении, например, заставляя модель неправильно определять дорожные знаки.

Предвзятость модели может крайне негативно повлиять на результат использования цифровых технологий. Среди возможных причин предвзятости – неравномерное распределение данных в обучающей выборке, алгоритмически заложенные предпочтения и предвзятое отношение к отдельным группам лиц. Даже классический спектр потенциальных кризисных кейсов огромен, начиная с эйджизма и сексизма при приеме на работу и заканчивая расизмом при выявлении потенциальных преступников [7].

Более того, недопустимо исключать ситуации, когда предвзятость модели является осмысленной политикой ее создателей, продиктованной их интересами, что несет еще большую опасность. Без должного контроля разработчики могут получить серьезную власть над обществом. При этом степень разрушительности последствий дискриминации трудно переоценить: систематические нарушения прав определенных социальных групп приведут к определению ИИ как бесчеловечного механизма.

Чтобы избежать проблем, которые могут нанести вред человеку путем искажения, хищения или утечки данных, необходимо убедиться, что результатам работы ИИ можно доверять. Так возникает потребность в понятии «доверенный искусственный интеллект»: система ИИ, в отношении которой пользователь может быть уверен, что она способна качественно выполнять возложенные на нее задачи [19].

2. Результаты и обсуждение: внедрение доверенного ИИ

Стоит признать, что внедрение искусственного интеллекта представляется весьма выгодным процессом. Однако именно доверие является ключевым фактором в вопросе применения ИИ, так как отторжение данной технологии массами исключит всю потенциальную пользу [5, 8, 13]. Следовательно, установление рациональных доверительных отношений, исключающих чрезмерное доверие или его отсутствие, позволит добиться выгоды для всего общества. Сущность этой идеи отражена в концепции доверенного ИИ.

Доверие к ИИ на физическом уровне означает уверенность в корректной работе всех его физических компонент, таких как датчики, и в качестве получаемых системой данных. Доверие к инфраструктуре, окружающей ИИ, означает уверенность в безопасности данных, с которыми ИИ взаимодействует, и в контроле над доступом к самой системе. Доверие на прикладном уровне означает уверенность в корректной работе ПО.

Если система ИИ считается доверенной, то есть доверие к ней проявляется на всех её уровнях, то такую систему можно допускать к решению задач с масштабными последствиями, так как пользователь может быть уверен в результатах ее работы.

Вопрос доверия достаточно комплексный и напрямую зависит от целого спектра особенностей человеческой психики [13, 20, 21]. Исследователи подчеркивают, что доверие есть желание и готовность отдельного лица зависеть от поступков другой стороны с целью извлечения некоторой выгоды, несмотря на потенциальные риски уязвимого положения. В основе этого феномена – совпадение нравственных ценностей, позволяющее предугадывать дальнейшие действия стороны, облеченной доверием, и уверенность в ее достаточной компетенции [20]. Также необходимо отметить, что кумулятивный характер доверия ставит дальнейшие отношения в зависимость от первого опыта. Значит, уже на изначальном этапе массового внедрения нейросетей должны быть четко выработаны стратегии выстраивания доверенного ИИ, отвечающего критериям этики и компетенции.

Далее представим несколько ключевых подходов, которые кажутся наиболее важными на пути внедрения доверенного искусственного интеллекта, для каждой из заинтересованных сторон.

А. Международное сообщество

Первый этап заключается в разработке законодательной базы, регулирующей работу ИИ в каждой из индустрий. Уже сейчас различные международные комитеты и организации заняты составлением общих рекомендаций и руководств к действию [21]. Несмотря на то, что отдельные попытки контролировать технологическую задачу предпринимались и ранее [22], отправной точкой в дискуссии вокруг этики и компетентности ИИ считается до-

кумент OЭСР Principles on Artificial Intelligence [23, 24]. Представив первичные наработки и ключевые принципы, авторы создали базу для будущих более серьезных документов [23, 25, 26]. Объединяя в себе наиболее значимые критерии доверенного ИИ, они нацелены на способствование выработке собственной политики странами – лидерами в области цифровых технологий [20].

Ключевые принципы включают постоянный человеческий контроль, надежность, подотчетность, конфиденциальность и прозрачность, но не ограничиваются ими. В будущих конвенциях могут также содержаться понятия непредвзятости и справедливости, сохранения социального и экологического благополучия, гибкости алгоритмов и введения так называемого права на объяснение [23, 25–27]. Важно подчеркнуть, что, пока органы власти будут обязаны не только постоянно адаптировать правовую базу, но и расширять ее, международные конвенции должны обеспечивать базисность объявленных принципов. В итоге, опираясь на данные принципы, власти смогут выработать сбалансированную систему норм, разграничив области регулирования между диспозитивными (soft laws) и императивными актами (hard laws) [7]. Важно осознавать созидательный характер процесса, когда вопрос контроля ИИ в госуправлении решается на основе уже имеющейся законодательной базы, а не вопреки ей. Таким образом, демократические государства смогут противодействовать риску нарушения прав человека, стоящих во главе угла и оберегаемых независимыми инстанциями и законом. Избегая даже потенциальной возможности предвзятого отношения как на этапе разработки, так и в процессе применения, государство предложит этические основания доверять ИИ. Однако, несмотря на краеугольное значение будущих норм, они лишь указывают на важность соблюдения закона в вопросах безопасности и качества, при этом оставаясь абстрактными правилами.

Концепция доверенного ИИ представлена на *рис. 1* и затрагивает множество аспектов.

B. Научное сообщество

Первостепенное значение имеет способность ученых и разработчиков защищать нейронные сети. В противном случае основные надежды на доверие нейросетям будут подорваны, и ситуация потенциально будет развиваться по сценарию смещения в сторону цифровой зависимости граждан от третьих сторон. Однако эволюция методов защиты происходит почти синхронно с аналогич-

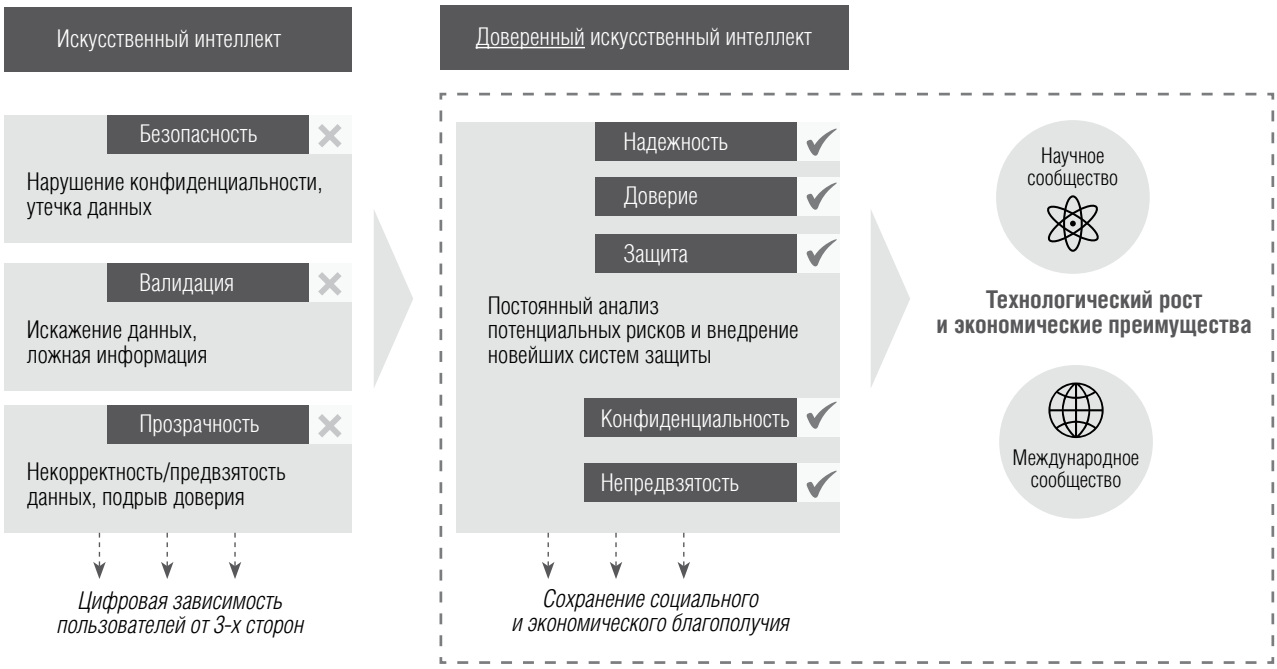

Рис. 1. Концепция использования доверенного ИИ.

ным процессом для атак на искусственный интеллект. Следовательно, в данном случае наиболее важно создание стандартов, которые гарантировали бы не только постоянный анализ потенциальных рисков, но и внедрение новейших систем защиты. Например:

1. Нарушение конфиденциальности (privacy breaches) – одна из наиболее вероятных проблем в будущем. Утечки личной информации могут возникнуть практически в любой момент работы нейронной сети – от обучения до вывода результатов. Очевидно, что данные частных лиц крайне выгодны определенным третьим лицам во многих сферах. В настоящий момент фокус разработчиков для защиты от потенциальных атак смещается с псевдонимизации и деидентификации к целому ряду новых технологий повышения конфиденциальности (PET). Одна из самых заметных – OPAL (open algorithms project) – заключается в предоставлении алгоритмам удаленного и контролируемого доступа к данным вместо пересылки анонимизированных данных. Таким образом, благодаря целому ряду специальных механизмов гарантируется возврат разработчикам модели только агрегированного результата, а запись всех операций обеспечивает возможность полной проверки процесса обучения [28, 29].

2. Другой сценарий, считающийся одним из наиболее опасных, – это отравление данных (*data poisoning*). Его идея состоит во включении в обучающую выборку ложной информации, либо обесценивающей результаты всей системы, либо подталкивающей нейросеть к принятию неверных решений, выгодных третьей стороне. Подобные технологии могут как потенциально нанести значительный урон репутации отдельной личности, так и повлиять на поведение масс (к примеру, на выборах) [20]. Конкретного решения пока не предложено. Однако, несмотря на весь свой угрожающий потенциал, риски в этой области пока минимальны: данные технологии все еще находятся на раннем этапе развития, не подразумевающем особых угроз. Тем не менее в литературе уже описано несколько моделей для борьбы с отравлением данных, почти каждая из которых акцентирует внимание на важности человеческого контроля и защиты.

3. Серьезное беспокойство вызывает и так называемая атака уклонения (evasion attack). В отличие от отравления данных, используемого во время фазы обучения, здесь угроза возникает на этапе применения ИИ. Незаметно видоизменяя входные значения, можно получить кардинально новый ответ сети. Даже малейшие преобразования могут привести к серьезным последствиям [7]. Имеется несколько механизмов противодействия атаке уклонения, однако самый широко описанный – это состязательная тренировка (adversarial training), подразумевающая включение разработчиками на этапе обучения заведомо неверных данных, позволяющих модели в дальнейшем не обращать внимания на потенциальные шумы [17, 30].

4. Очевидна интеллектуальная ценность уже обученной нейронной сети, особенно при использовании собираемых государством больших данных. Инверсия модели (model extraction) может потенциально означать колоссальные утечки персональных данных. Подобное нарушение конфиденциальности способно привести к крайне тяжелым последствиям, например, при утечке медицинских карт [7]. Так как сама технология существует пока только в научных статьях и абстрактных моделях, методы борьбы с ней тоже представлены лишь в теории. Среди наиболее признанных можно выделить приватное агрегирование обучающих моделей (PETA). Идея данной концепции состоит в разъединении данных на несколько наборов – «учителей», каждый из которых обучает отдельную нейросеть. Далее эти независимые модели объединяются для обучения нейросети «ученик» путем голосования, в итоге, не давая последней доступа к изначальным данным [18, 31].

5. В научной литературе представлен целый ряд метрик, позволяющих с определенной эффективностью обнаружить предвзятости модели (difference in means, difference in residuals, equal opportunity, equal mis-opportunity, disparate impact, normalized mutual information). Опираясь на данные математические подходы, разработчики смогут выявлять несовершенства своих нейросетей и вносить соответствующие изменения. Смягчение предвзятости может быть достигнуто за счет одной из трех категорий алгоритмов: pre-processing algorightms, in-processing algorithms и post-processing algorithms [32].

Наконец, важным аспектом политики объединенного научного сообщества должна оказаться своевременность представления конкретных технологий, которые позволят более эффективно реализовывать заявленные принципы. Даже сейчас разрыв между используемыми алгоритмами и абстрактными идеалистическими теориями очевиден. В то же время уже давно проводятся многочисленные углубленные исследования, изучающие возможные сценарии и инструменты.

Несколько подходов к решению вышеуказанных проблем уже были представлены. Один из них побуждает нас комплексно разобрать технологию блокчейн – непрерывную цепочку блоков, связанных в обратном порядке посредством хеш-сумм. Каждый блок, помимо собственного хеша и хеша предыдущего блока, содержит некоторую информацию. Таким образом, блокчейн представляет собой распределенный реестр общего пользования, позволяющий хранить и передавать данные с особо высокой надежностью и без возможности вмешаться. Интеллектуальные контракты (smart contracts), то есть программный код, обеспечивающий выполнение всех установленных условий, в связке с машинным обучением гарантирует надежность итоговых результатов.

Следовательно, объединение технологии блокчейн с ИИ позволит создать децентрализованную систему, хранящую информацию любой ценности и предоставляющую ее для обучения нейросети. В итоге не только гарантируется безопасность данных, но и решается проблема черного ящика благодаря полной и при этом децентрализованной истории операций, исключающей внешнее вмешательство или определенную предвзятость [24, 33].

Также появляются инструменты для отслеживания прозрачности и справедливости моделей. Компания IBM представила несколько инструментов для разработки доверенного искусственного интеллекта: библиотеку на Python, помогающую проверять модели на устойчивость к различным атакам – Adversarial Robustness Toolbox (ART); алгоритмы, анализирующие процесс работы моделей ИИ для создания прозрачности – AI Explainability 360; библиотеку с метриками, показывающими справедливость модели – AI Fairness 360 [34].

Выводы

Повсеместная интеграция цифровых технологий и создание децентрализованных экосистем способны открыть новые горизонты и устранить ряд текущих проблем. Ключевая роль нейронных сетей в этом процессе заставляет уделять особое внимание каждой потенциальной угрозе. Особого внимания заслуживает вопрос возникновения цифровой защиты и использование доверенного искусственного интеллекта. Совместный контроль за соблюдением принципов доверенного ИИ позволит как научному сообществу, так и международным организациям создать и регулярно обновлять механизмы противодействия рискам. Более того, у человечества в целом появится шанс грамотно распорядиться своим политическим опытом, накопленным за столетия.

Несмотря на то, что почти каждая проблема имеет конкретное решение, представленные технологические стратегии — не более чем теоретизация на тему планирующихся в ближайшем будущем процессов. Другими словами, идеалистические тезисы, основывающиеся на биоэтике, будут неизменно сталкиваться с реальностью, обусловленной ситуацией в каждой отдельной стране. Именно поэтому важно постоянно адаптировать абстрактные принципы под изменяющуюся реальность, расширять законодательную базу и находить новые решения. В то же время крайне необходима твердая коллективная приверженность сохранению демократических институтов.

Научный прогресс уже сейчас позволяет делать первые шаги в сторону массового внедрения ИИ. Хотя нынешние масштабы относительно скромны, перспективы вдохновляют. Если обществу удастся достойно справиться с предстоящими вызовами, перечисленные выше изменения поднимут человечество на совершенно новый уровень существования. ■

Литература

1. Vogt T., Winter P., Nessler B., Doms T. Trusted artificial intelligence: towards certification of machine learning applications. Vienna: TÜV Austria Holding AG, 2021.
2. Kuleshov A., Ignatiev A., Abramova A., Marshalko G. Addressing AI ethics through codification // 2020 International Conference Engineering Technologies and Computer Science (EnT), 2020. P. 24–30. https://doi.org/10.1109/EnT48576.2020.00011
3. Harrison T., Luna-Reyes L. Cultivating trustworthy artificial intelligence in digital government // Social Science Computer Review. 2021. Vol. 40. No. 2. P. 494–511. https://doi.org/10.1177/0894439320980122
4. Recommendation of the council on digital government strategies. Paris: OECD Publishing, 2014. [Электронный ресурс]: https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf (дата обращения: 22.09.2021).
5. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. Adversarial attacks and defences: A survey. Working paper arXiv: 1810.00069, 2018. https://doi.org/10.48550/arXiv.1810.00069

6. Haenlein M., Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence // California Management Review. 2019. Vol. 61. No. 4. P. 5–14. https://doi.org/10.1177/0008125619864925

7. Wei J. Research progress and application of computer artificial intelligence technology // MATEC Web of Conferences. 2018. Vol. 176. Article Number 01043. https://doi.org/10.1051/matecconf/201817601043

8. Mijwil M., Esen A., Alsaadi A. Overview of neural networks. 2019. [Электронный ресурс]: https://www.researchgate.net/publication/332655457_Overview_of_Neural_Networks (дата обращения: 22.09.2021).

9. Sibai F.N. AI crimes: A classification // 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). 15–19 June 2020. P. 1–8. https://doi.org/10.1109/CyberSecurity49315.2020.9138891

10. Jastroch N. Trusted artificial intelligence: On the use of private data // Product Lifecycle Management Enabling Smart X. PLM 2020 (eds. F. Nyffenegger, J. Ríos, L. Rivest, A. Bouras). IFIP Advances in Information and Communication Technology. 2020. Vol. 594. https://doi.org/10.1007/978-3-030-62807-9_52

11. Nemitz P. Constitutional democracy and technology in the age of artificial intelligence // Phil. Trans. R. Soc. A. 2018. Vol. 376. http://doi.org/10.1098/rsta.2018.0089

12. Misra S.K., Das S., Gupta S., Sharma S.K. Public policy and regulatory challenges of artificial intelligence (AI) // Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation. TDIT 2020 (eds. S.K. Sharma, Y.K. Dwivedi, B. Metri, N.P. Rana). IFIP Advances in Information and Communication Technology. 2020. Vol. 617. https://doi.org/10.1007/978-3-030-64849-7_10

13. Павлютенкова М.Ю. Электронное правительство vs цифровое правительство в контексте цифровой трансформации // Мониторинг общественного мнения: Экономические и социальные перемены. 2019. № 5. С. 120–135. https://doi.org/10.14515/monitoring.2019.5.07

14. Концепция цифрового правительства как политический проект для России: перспективы реализации в условиях вызовов и рисков цифровизации общества / А.П. Кочетков и [др.] // Власть. 2021. Т. 29. № 1. С. 317–331. https://doi.org/10.31171/vlast.v29i1.7963

15. Williams M., Valayer C. Digital government benchmark – study on digital government transformation. European Union, 2018. [Электронный ресурс]: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-digital-government-benchmark-study-digital-government-transformation (дата обращения: 22.09.2021).

16. Carter D. Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? // Business Information Review. 2020. Vol. 37. No. 2. P. 60–68. https://doi.org/10.1177/0266382120923962

17. Kamat G. Algorithms for private data analysis. Lecture 14 – Private ML and stats: Modern ML. [Электронный ресурс]: http://www.gautamkamath.com/CS860notes/lec14.pdf (дата обращения: 02.09.2021).

18. Hinnefeld J., Cooman P., Mammo N., Deese R. Evaluating fairness metrics in the presence of dataset bias. Working paper arXiv: 1809.09245, 2018. https://doi.org/10.48550/arXiv.1809.09245

19. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. ГОСТ Р 59276-2020 // Национальный стандарт Российской Федерации, 2021.

20. Tinholt D., Carrara W., Linden N. Unleashing the potential of Artificial Intelligence in the Public Sector. Capgemini, 2017. [Электронный ресурс]: https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/10/ai-in-public-sector.pdf (дата обращения: 22.09.2021).

21. Sharma G.D., Yadav A., Chopra R. Artificial intelligence and effective governance: A review, critique and research agenda // Sustainable Futures. 2020. Vol. 2. Article ID 100004. https://doi.org/10.1016/j.sftr.2019.100004

22. McCormick T.R., Min D. Principles of Bioethics. University of Washington, 2020. [Электронный ресурс]: https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/articles/principles-bioethics (дата обращения: 22.09.2021).

23. Lindgren I., Veenstra A.F. Digital government transformation: a case illustrating public e-service development as part of public sector transformation // Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands, 2018. P. 1–6.

24. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, Version 2. The IEEE global initiative on ethics of autonomous and intelligent systems. IEEE, 2021. [Электронный ресурс]: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (дата обращения: 22.09.2021).

25. Брадул Н.В., Лебезова Э.М. Концептуализация понятия «Smart Government»: наукометрический подход // Управленец. 2020. Т. 11. № 3. С. 33–45. https://doi.org/10.29141/2218-5003-2020-11-3-3

26. Thiebes S., Lins S., Sunyaev A. Trustworthy artificial intelligence // Electron Markets. 2021. Vol. 31. P. 447–464. https://doi.org/10.1007/s12525-020-00441-4

27. Falco G., Viswanathan A., Caldera C., Shrobe H. A master attack methodology for an AI-based automated attack planner for smart cities // IEEE Access. 2018. Vol. 6. P. 48360-48373. https://doi.org/10.1109/ACCESS.2018.2867556.

28. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias / R. Bellamy [et al.] Working paper arXiv: 1810.01943, 2018. https://doi.org/10.48550/arXiv.1810.01943

29. Towards enabling trusted artificial intelligence via Blockchain / K. Sarpatwar [et al.] // Calo, S., Bertino, E., Verma, D. (eds) Policy-Based Autonomic Data Governance. Lecture Notes in Computer Science. Vol. 11550. P. 137–153. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-17277-0_8

30. Montjoye Y.D., Farzanehfar A., Hendrickx J., Rocher L. Solving artificial intelligence's privacy problem // Field Actions Science Reports. 2017. Special Issue 17. P. 80–83. [Электронный ресурс]: https://journals.openedition.org/factsreports/pdf/4494 (дата обращения: 22.09.2021).

31. Open Algorithms (OPAL) Project. DataCollaboratives.org, 2021. [Электронный ресурс]: https://datacollaboratives.org/cases/open-algorithms-opal-project.html (дата обращения: 22.09.2021).

32. Salah K., Rehman M.H.U., Nizamuddin N., Al-Fuqaha A. Blockchain for AI: Review and open research challenges // IEEE Access. 2019. Vol. 7. P. 10127–10149. https://doi.org/10.1109/ACCESS.2018.2890507

33. Baker-Brunnbauer J. Management perspective of ethics in artificial intelligence // AI and Ethics. 2021. Vol. 1. P. 173–181. https://doi.org/10.1007/s43681-020-00022-3

34. IBM. [Электронный ресурс]: https://www.ibm.com/us-en/ (дата обращения: 22.09.2021).

Об авторах

Авдошин Сергей Михайлович

кандидат технических наук;

профессор департамента компьютерной инженерии, Московский институт электроники и математики им. А.Н. Тихонова, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

Песоцкая Елена Юрьевна

кандидат экономических наук;

доцент департамента программной инженерии, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645

Trusted artificial intelligence: Strengthening digital protection

Sergey M. Avdoshin

E-mail: savdoshin@hse.ru

Elena Yu. Pesotskaya

E-mail: epesotskaya@hse.ru

HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This article is devoted to aspects associated with the up-coming need for mass implementation of neural networks in the modern society. On the one hand, the latter will fully expand the capabilities of state institutions and society delegated to perform numerous tasks with higher efficiency. However, a significant threat to democratic institutions obliges society to set out the concept of reliable artificial intelligence (AI). The authors explore a new concept of a trusted AI necessary for the scientific and international community to counter improper future digital penetration. Explaining to what extent digital transformation is mandatory, the authors emphasize the numerous dangers

associated with the applications of artificial intelligence. The purpose of the article is to study the potential hazards of neural networks' abuse by the authorities and the resistance to them with reliance on the trusted AI. Studying various aspects of digital transformation and the use of artificial intelligence technologies, the authors formalize the dangers associated with the emergence and propose an approach to the use of digital protection technologies that can be trusted.

Keywords: neural networks, digital protection, trust, artificial intelligence, society

Citation: Avdoshin S.M., Pesotskaya E.Yu. (2022) Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, vol. 16, no. 2, pp. 62–73. DOI: 10.17323/2587-814X.2022.2.62.73

References

1. Vogt T., Winter P., Nessler B., Doms T. (2021) *Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications*. Vienna: TÜV Austria Holding AG.
2. Kuleshov A., Ignatiev A., Abramova A., Marshalko G. (2020) Addressing AI ethics through codification. *2020 International Conference Engineering Technologies and Computer Science (EnT),* pp. 24–30. https://doi.org/10.1109/EnT48576.2020.00011
3. Harrison T., Luna-Reyes L. (2021) Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review*, vol. 40, no. 2, pp. 494–511. https://doi.org/10.1177/0894439320980122
4. OECD (2014) *Recommendation of the council on digital government strategies.* Paris: OECD Publishing. Available at: https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf (accessed 22 September 2021).
5. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. (2018) *Adversarial attacks and defences: A survey*. Working paper arXiv: 1810.00069. https://doi.org/10.48550/arXiv.1810.00069
6. Haenlein M., Kaplan A. (2019) A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, vol. 61, no. 4, pp. 5–14. https://doi.org/10.1177/0008125619864925
7. Wei J. (2018) Research progress and application of computer artificial intelligence technology. *MATEC Web of Conferences*, vol. 176, Article Number 01043. https://doi.org/10.1051/matecconf/201817601043
8. Mijwil M., Esen A., Alsaadi A. (2019) *Overview of neural networks*. Available at: https://www.researchgate.net/publication/332655457_Overview_of_Neural_Networks (accessed 22 September 2021).
9. Sibai F.N. (2020) AI crimes: A classification. *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pp. 1–8. https://doi.org/10.1109/CyberSecurity49315.2020.9138891
10. Jastroch N. (2020) Trusted artificial intelligence: On the use of private data. In: *Product Lifecycle Management Enabling Smart X. PLM 2020* (eds. F. Nyffenegger, J. Ríos, L. Rivest, A. Bouras). IFIP Advances in Information and Communication Technology, vol. 594. https://doi.org/10.1007/978-3-030-62807-9_52
11. Nemitz P. (2018) Constitutional democracy and technology in the age of artificial intelligence. *Phil. Trans. R. Soc. A.*, vol. 376. http://doi.org/10.1098/rsta.2018.0089
12. Misra S.K., Das S., Gupta S., Sharma S.K. (2020) Public Policy and Regulatory Challenges of Artificial Intelligence (AI). In: *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation. TDIT 2020* (eds. S.K. Sharma, Y.K. Dwivedi, B. Metri, N.P. Rana). IFIP Advances in Information and Communication Technology, vol. 617. https://doi.org/10.1007/978-3-030-64849-7_10
13. Pavlutenkova M. (2019) Electronic government vs digital government in context of digital transformation. *Monitoring of Public Opinion: Economic and Social Changes Journal*, no. 5, pp. 120—135 (in Russian). https://doi.org/10.14515/monitoring.2019.5.07
14. Kochetkov A.P., Vasilenko I.A., Volodenkov S.V., Gadzhiev K.S., Kovalenko V.I., Soloviev A.I., Kirsanova E.G. (2021) Political Project for Russia: Prospects for implementation in the context of challenges and risks of digitalization of society. *Vlast' (The Authority)*, vol. 29, no. 1, pp. 317—331 (in Russian). https://doi.org/10.31171/vlast.v29i1.7963
15. Williams M., Valayer C. (2018) *Digital government benchmark – study on digital government transformation*. European Union. Available at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-digital-government-benchmark-study-digital-government-transformation (accessed 22 September 2021).
16. Carter D. (2020) Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review*, vol. 37, no. 2. pp. 60–68. https://doi.org/10.1177/0266382120923962
17. Kamat G. (2014) *Algorithms for private data analysis. Lecture 14 – Private ML and stats: Modern ML*. Available at: http://www.gautamkamath.com/CS860notes/lec14.pdf (accessed 22 September 2021).
18. Hinnefeld J., Cooman P., Mammo N., Deese R. (2018) *Evaluating fairness metrics in the presence of dataset bias*. Working paper arXiv: 1809.09245. https://doi.org/10.48550/arXiv.1809.09245

19. National Standard of the Russian Federation (2021) *Artificial intelligence systems. Methods for ensuring trust. General. GOST R 59276-2020* (in Russian).

20. Tinholt D., Carrara W., Linden N. (2017) *Unleashing the potential of Artificial Intelligence in the Public Sector*. Capgemini. Available at: https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/10/ai-in-public-sector.pdf (accessed 22 September 2021).

21. Sharma G.D., Yadav A., Chopra R. (2020) Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, vol. 2, Article ID 100004. https://doi.org/10.1016/j.sftr.2019.100004

22. McCormick T.R., Min D. (2020) *Principles of Bioethics*. University of Washington. Available at: https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/articles/principles-bioethics (accessed 22 September 2021).

23. Lindgren I., Veenstra A.F. (2018) Digital government transformation: a case illustrating public e-service development as part of public sector transformation. Proceedings of the *19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands*, pp. 1–6.

24. IEEE (2021) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2.* The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed 22 September 2021).

25. Bradul N.V., Lebezova E.M. (2020) Conceptualization of Smart Government: A scientometric approach. *Upravlenets (The Manager)*, vol. 11, no. 3, pp. 33–45. https://doi.org/10.29141/2218-5003-2020-11-3-3 (in Russian).

26. Thiebes S., Lins S., Sunyaev A. (2021) Trustworthy artificial intelligence. *Electron Markets*, vol. 31, pp. 447–464. https://doi.org/10.1007/s12525-020-00441-4

27. Falco G., Viswanathan A., Caldera C., Shrobe H. (2018) A master attack methodology for an AI-based automated attack planner for smart cities. *IEEE Access*, vol. 6, pp. 48360-48373. https://doi.org/ 10.1109/ACCESS.2018.2867556.

28. Bellamy R. et al. (2018) *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* Working paper arXiv: 1810.01943. https://doi.org/10.48550/arXiv.1810.01943

29. Sarpatwar K. et al. (2019) Towards enabling trusted artificial intelligence via Blockchain. In: Calo, S., Bertino, E., Verma, D. (eds) *Policy-Based Autonomic Data Governance. Lecture Notes in Computer Science*, vol. 11550, pp. 137–153. Springer, Cham. https://doi.org/10.1007/978-3-030-17277-0_8

30. Montjoye Y.D., Farzanehfar A., Hendrickx J., Rocher L. (2017) Solving artificial intelligence's privacy problem. *Field Actions Science Reports*, Special Issue 17, pp. 80—83. Available at: https://journals.openedition.org/factsreports/pdf/4494 (accessed 22 September 2021).

31. DataCollaboratives.org (2021) *Open Algorithms (OPAL) Project*. Available at: https://datacollaboratives.org/cases/open-algorithms-opal-project.html (accessed 22 September 2021).

32. Salah K., Rehman M.H.U., Nizamuddin N., Al-Fuqaha A. (2019) Blockchain for AI: Review and open research challenges. *IEEE Access*, 2019, vol. 7, pp. 10127–10149. https://doi.org/10.1109/ACCESS.2018.2890507

33. Baker-Brunnbauer J. (2021) Management perspective of ethics in artificial intelligence. *AI and Ethics*, vol. 1, pp. 173–181. https://doi.org/10.1007/s43681-020-00022-3

About the authors

Sergey M. Avdoshin

Cand. Sci. (Tech.);

Professor, School of Computer Engineering, HSE Tikhonov Moscow Institute of Electronics and Mathematics (MIEM HSE), National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

Elena Yu. Pesotskaya

Cand. Sci. (Econ.);

Associate Professor, School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645