УДК 331.101 DOI 10.26425/1816-4277-2024-3-176-185

Проблемы предвзятости и дискриминации человеческого капитала в системах искусственного интеллекта

Каштанова Екатерина Викторовна

Канд. экон. наук, доц. каф. управления персоналом
ORCID: 0000-0002-5949-9198, e-mail: kashtanovae@mail.ru

Лобачева Анастасия Сергеевна

Канд. экон. наук, доц. каф. управления персоналом
ORCID: 0000-0002-4210-9018, e-mail: aslobacheva@mail.ru

Государственный университет управления, г. Москва, Россия

Аннотация

Статья посвящена изучению генезиса понимания причин предвзятости и дискриминации человеческого капитала от традиционных форм до эпохи цифровизации и искусственного интеллекта (далее – ИИ). В данной статье авторы продолжают свои тематические исследования в области изучения вопросов этики ИИ, преимуществ и рисков его повсеместного распространения и использования. Основной целью написания этой работы является изучение вопроса о том, насколько усугубляет применение ИИ и алгоритмизированных решений проблему предвзятости и дискриминации человеческого капитала, которая неизбежно связана с появлением ИИ. Авторы статьи представляют краткий обзор ретроспективы, а также открывают новые современные формы проявления дискриминации, порождаемые действием ИИ, и подвергают их открытому обсуждению, предлагая свое видение решения вопросов по нейтрализации рисков, распространяемых применением технологий ИИ в отношении тех или иных групп трудящихся. Авторы рассматривают в статье проявления предвзятости и дискриминации в обществе в целом и в сфере управления человеческими ресурсами в частности, определяют возможные угрозы дискриминации в результате распространения ИИ и последствия этих угроз.

Ключевые слова

Искусственный интеллект, дискриминация человеческого капитала, цифровизация, цифровая грамотность, цифровая компетентность, предвзятость искусственного интеллекта, этика искусственного интеллекта, цифровая предвзятость

Для цитирования: Каштанова Е.В., Лобачева А.С. Проблемы предвзятости и дискриминации человеческого капитала в системах искусственного интеллекта//Вестник университета. 2024. № 3. С. 176–185.

Problems of bias and discrimination of human capital in artificial intelligence systems

Ekaterina V. Kashtanova
Cand. Sci. (Econ.), Assoc. Prof at the Personnel Management Department
ORCID: 0000-0002-5949-9198, e-mail: kashtanovae@mail.ru

Anastasia S. Lobacheva
Cand. Sci. (Econ.), Assoc. Prof at the Personnel Management Department
ORCID: 0000-0002-4210-9018, e-mail: aslobacheva@mail.ru

State University of Management, Moscow, Russia

Abstract

The article is devoted to the study of the genesis of understanding the causes of bias and discrimination of human capital from traditional forms to the era of digitalisation and artificial intelligence (hereinafter referred to as AI). In this article, the authors continue their case studies in the field of studying the ethics of AI, advantages and risks of its widespread distribution and use. The main purpose of writing this work is to study the question of how the use of AI and algorithmic solutions aggravates the problem of bias and discrimination of human capital which is inevitably associated with the emergence of AI. The authors of the article present a brief overview of the retrospective, discover new modern forms of discrimination generated by the action of AI and subject them to open discussion, offering their vision of solving issues on neutralisation the risks caused by the use of AI technologies in relation to certain groups of workers. The authors consider in the article the manifestations of bias and discrimination in society in general and in the field of human resource management in particular, identify possible threats of discrimination as a result of the spread of AI and the consequences of these threats.

Keywords

Artificial intelligence, human capital discrimination, digitalisation, digital literacy, digital competence, artificial intelligence bias, artificial intelligence ethics, digital bias

For citation: Kashtanova E.V., Lobacheva A.S. (2024) Problems of bias and discrimination of human capital in artificial intelligence systems. *Vestnik universiteta*, no. 3, pp. 176–185.

ВВЕДЕНИЕ

В последнее время не только работодатели или ученые, но и каждый отдельный человек все больше ощущает на себе действие технологий искусственного интеллекта (далее – ИИ) в самых разных областях жизни и практической деятельности. Зададимся вопросом, порождает ли применение ИИ и цифровых алгоритмизированных решений проблему дискриминации человеческого капитала. Под ИИ мы понимаем способность компьютера обучаться, принимать решения и совершать действия, свойственные человеческому интеллекту [1]. Цифровые алгоритмизированные решения представляют собой программы, по которым действует ИИ.

Действительно, ИИ все более глубоко проникает не только в сферу управления организацией, но и в мир в целом, влияя на принятие жизненно важных решений, таких как трудоустройство, оценка персонала или доступные социальные льготы. Это повышает риск дискриминации человеческого капитала со стороны ИИ. Путь к управлению и снижению этого риска начинается с понимания того, как может возникнуть такая дискриминация и почему ее трудно обнаружить.

В ближайшей перспективе цель сохранения благотворного влияния ИИ на общество мотивирует исследования во многих областях, от экономики и права до технических тем, даже таких, как безопасность и контроль [2]. Незначительное мошенничество или неявная несправедливость в киберпространстве пока является совершенно ничтожным (в какой-то степени даже побочным) действием по сравнению с глобальным преимуществом, которое дает сегодня система ИИ – учиться делать то, что хочет от нее человек, принимая на себя его традиционные, чаще рутинные, функции.

В долгосрочной же перспективе главнейший вопрос состоит в том, а что случится, если новый мощный ИИ станет гораздо лучше и эффективнее людей решать все их задачи? Однако многие эксперты уже сейчас высказывают опасения по поводу такого развития событий и заявляют о том, что в случае, если не научиться согласовывать действия ИИ, то власть человека на земле на этом может закончиться. Мы считаем, что необходимо сформировать понимание важности данного вопроса и привлечь к нему пристальное внимание всех заинтересованных лиц.

ДИСКРИМИНАЦИЯ В ОБЩЕСТВЕ: ПРИЧИНЫ И ПРИМЕРЫ ПРОЯВЛЕНИЯ

Вопросы дискриминации со стороны ИИ тесно связаны с ведущейся в научных кругах дискуссией об этике ИИ. Например, есть мнение, что алгоритмы ИИ наносят вред системе социальной защиты, криминализируют бедных, усиливают дискриминацию и ставят под угрозу национальные ценности [3]. По мнению других авторов, если все общество не будет соблюдать этику и требовать обеспечения справедливости в процессе обмена информацией и передачи данных, дискриминация, вызванная ИИ, будет продолжать расти [4]. Мы можем согласиться с мнением, что нынешние причины дискриминации, вызванные действием ИИ, исходят от тех же концептуальных проблем, которые были характерны дискриминации с самого начала ее формального толкования в законе и этике [5].

Понятие дискриминации широко используется в повседневной речи, а также во многих национальных законодательствах и наднациональных кодексах. Достаточно трудно истолковать суть дискриминации таким образом, чтобы учесть все значения или, по крайней мере, большинство значений данного понятия. Для того чтобы наилучшим образом выразить смысл дискриминации человеческого капитала, которая может возникнуть (и уже возникает) с приходом и распространением ИИ, обратимся к ее проявлениям, которые существовали до прихода ИИ.

Ярким примером дискриминации выступает разделение населения на касты в Индии. Одной из причин, по которой кастовая система смогла существовать, была функциональная взаимозависимость в индуистском обществе, позволяющая проводить широкий спектр дифференциации, не нарушая социальную структуру населения. Религиозная основа, заложенная в кастовую систему, давала высшим кастам возможность увековечивать различия и пользоваться растущими привилегиями для подавления низших каст.

В медицине только в конце 1950-х гг. дискриминация стала основной темой исследований. До этого времени люди с отклонениями от установленных обществом поведенческих норм считались больными с патологией, а само общество исключало возможности для таких людей в получении образования, нормальной жизни, признания. Во многих странах мира инвалиды были отстранены от участия в общественных делах из-за физических проблем со здоровьем, препятствующих их мобильности, а предвзятость и дискриминация по отношению к ним в лучшем случае затрудняли им учебу и работу.

Можно констатировать тот факт, что и сегодня симптомы негативного поведения распространены по отношению к людям с какими-то физическими особенностями (например, с чересчур избыточным весом), отклоняющимися от нормы, принятой в конкретном обществе. В Новом Завете упоминается, что такая болезнь, как проказа, также считалась определяющим фактором дискриминации. Больной способен излечиться исключительно с помощью божественного чуда, и даже сегодня изоляция больных людей продолжает осуществляться как наиболее результативная стратегия борьбы с распространением заболеваний во время пандемии. Кроме того, наиболее известным проявлением является дискриминация людей по расовому, классовому и гендерному признакам.

Таким образом, нормативные установки, как культурные, так и межличностные, явным образом оказывают влияние на поведение большинства в силу традиций, мировоззрения, системы ценностей и убеждений, а также культивируемого в обществе эталона человека (его внешность, манеры поведения, религиозные воззрения, род деятельности) и выступают причиной появления дискриминации.

ПРЕДВЗЯТОСТЬ ИИ И ДИСКРИМИНАЦИЯ ЧЕЛОВЕЧЕСКОГО КАПИТАЛА

Дискриминация человеческого капитала определяется как ситуация, в которой к работникам или группам работников относятся по-разному с точки зрения найма, оплаты труда, льгот и продвижения по службе по сравнению с другими работниками или группами из-за их неэкономических характеристик, включая пол, расу, религию и возраст. Человеческий капитал – это актив, состоящий из знаний и навыков, которыми обладает человек и которые могут быть использованы организацией для достижения своих целей. Человеческий капитал важен, потому что определенный уровень человеческих знаний и навыков необходим для того, чтобы организация могла достигать данных целей.

Дискриминация человеческого капитала тесно связана с принадлежностью человека к определенной группе, однако, как известно, ни один вид группового членства не имеет права на проявление какой-либо предвзятости по отношению к другому виду. Правовое положение о дискриминации отражено в статье 26 Международного пакта о гражданских и политических правах.

«Все люди равны перед законом и имеют право на равную защиту закона без любых проявлений дискриминации. Справедливый закон запрещает любую дискриминацию и обеспечивает гарантию равноправия для всех, также закон должен обеспечивать эффективную защиту от дискриминации по любому основанию, таким как раса, цвет кожи, пол, язык, религия, политические или иные взгляды, национальное или социальное происхождение, имущество, рождение или другой статус»[1].

Казалось бы, что все вопросы, связанные с дискриминацией, изучены и обрели правовой статус. Однако на сегодняшний момент возникает новое проявление предвзятости, которое весьма трудно предугадать и предусмотреть. Эта новая волна дискриминации связана с распространением и проникновением во все сферы жизни ИИ. ИИ действительно начинает технологическую революцию, и, хотя ему еще предстоит «захватить» мир, есть более насущная проблема, с которой мы уже столкнулись, – предвзятость ИИ.

Предвзятость ИИ – это основное предубеждение в данных, которые используются для создания алгоритмов ИИ, что в конечном итоге может привести к дискриминации и другим последствиям [6]. Приведем простой пример. Представим, что мы хотим создать алгоритм, который решает, будет ли абитуриент принят в университет или нет, и одним из наших входных данных будет географическое местоположение абитуриента. Гипотетически, если бы местоположение человека сильно коррелировало с этнической принадлежностью, то наш алгоритм косвенно отдавал бы предпочтение определенным этническим группам перед другими. Это пример предвзятости в ИИ.

ПРЕДВЗЯТОСТЬ И ДИСКРИМИНАЦИЯ ИИ: ПРИМЕРЫ ИЗ ПРАКТИКИ

Приведем реальные примеры того, когда алгоритмы ИИ демонстрировали предубеждение и дискриминацию по отношению к людям.

В октябре 2019 г. исследователи обнаружили, что алгоритм, практикуемый более чем на 200 млн чел. в больницах Соединенных Штатов Америки для прогнозирования того, какие пациенты, вероятно, будут нуждаться в дополнительной медицинской помощи, в значительной степени отдавал предпочтение белокожим пациентам по сравнению с чернокожими пациентами. Хотя сама раса не была переменной, заложенной

[1] Организация Объединенных Наций. Международный пакт о гражданских и политических правах. Режим доступа: https://www.un.org/ru/documents/decl_conv/conventions/pactpol.shtml (дата обращения: 20.01.2024).

в этом алгоритме, другой переменной, сильно коррелирующей с расой, была история затрат на здравоохранение. Обоснование состояло в том, что стоимость суммирует количество потребностей в медицинском обслуживании у конкретного человека. По разным причинам темнокожие пациенты в среднем несли более низкие расходы на медицинское обслуживание, чем белокожие пациенты с теми же заболеваниями.

Приведем другой пример. Amazon – один из крупнейших технологических гигантов в мире. Поэтому неудивительно, что в компании активно используется цифровизация процессов управления персоналом и ИИ. В 2015 г. представители компании осознали, что их алгоритм, используемый для найма сотрудников, оказался предвзятым в отношении женщин. Причина этого заключалась в том, что алгоритм основывался на количестве резюме, поданных за последние 10 лет, и, поскольку большинство заявителей были мужчинами, ИИ стал отдавать предпочтение мужчинам, а не женщинам.

Цифровизация, которую представляют обществу под лозунгом удобства, – это гражданская цифровизация. Внутри общества распространение ИИ в первую очередь влечет полную ликвидацию приватности, что является огромной возможностью для дискриминации человеческого капитала. Самым ярким примером дискриминации человеческого капитала в эпоху ИИ выступает угроза внедрения системы социальных рейтингов. Система социальных рейтингов – это система оценки, в основу которой положено социально-политическое поведение индивидов, организаций и других учреждений для определения их социальной репутации. На ее основе реализуется политика стимулирования и санкций в отношении них.

Как утверждает И.С. Ашманов, член совета при Президенте Российской Федерации (далее – РФ, Россия) по развитию гражданского общества и правам человека, предприниматель в сфере информационных технологий (далее – ИТ) и ИИ, им в DarkNet (англ. теневой интернет) в качестве эксперимента за два часа была приобретена полная информация о человеке, включая его банковские счета, активы, данные паспорта, образование, место жительства и работы. Самое интересное, что вместе с этими данными продавалось «передвижение по городу в течение дня», то есть видеопуть человека с камер системы «Безопасный город» с передачей изображения с камеры на камеру. Вся эта информация стоила менее 10 тыс. руб.[2]

Кроме этого, сегодня человек уступил большую часть решений сложным машинам. Автоматическое право на участие системы в процессе принятия решений, алгоритмы ранжирования и модели прогнозирования риска контролируют и определяют, какие семьи получают необходимые дотации, кто попадает в шорт-лист для трудоустройства и кто может быть наиболее склонен к мошенничеству. Известны случаи, когда система отказывала в доступе, например, в транспорт или торговые центры людям с определенным типом внешности, потому что системы безопасности, основанные на ИИ, определяли эту внешность как потенциально опасную (например, сходство с образом террориста).

ДИСКРИМИНАЦИОННЫЕ ЭФФЕКТЫ ИИ В СФЕРЕ УПРАВЛЕНИЯ ЧЕЛОВЕЧЕСКИМИ РЕСУРСАМИ

Далее мы покажем, в каких областях управления человеческими ресурсами используемые алгоритмы принятия решений и другие типы ИИ создают дискриминационные эффекты или могут создавать их в обозримом будущем.

Принятие решений на основе ИИ может привести к дискриминации несколькими способами. Одним из них является определение целевой переменной и меток классов. Именно в вопросах управления человеческими ресурсами наиболее часто возникает ситуация выбора, начиная от отбора персонала на работу и заканчивая вопросами продвижения по должностям и увольнения. Предположим, компании нужна система ИИ для сортировки заявлений о приеме на работу, чтобы найти хороших сотрудников. Однако как следует определять хорошего сотрудника? Другими словами, какими должны быть метки класса хороших сотрудников? Хороший сотрудник – тот, кто продает больше всего товаров или тот, кто никогда не опаздывает на работу? Тем не менее, тогда кандидат на должность, живущий дальше от месторасположения компании, будет определен ею как потенциально опаздывающий.

Также исследования применимости ИИ в области управления человеческими ресурсами приводят к неожиданным результатам. Интересным оказался факт, что некоторые менеджеры неохотно соглашаются

[2] Ашманов И.С. Чудовищная антиутопия в реальности. Режим доступа: http://новости-россии.ru-an.info/новости/началась-чудовищная-антиутопия-в-реальности-или-искусственный-интеллект-наступает/ (дата обращения: 21.01.2024).

с ИИ, опасаясь, что он может дискриминировать их рабочие роли, их значение как руководителей и сократить их влияние на рабочем месте. Эти менеджеры склонны интерпретировать ИИ как угрозу для карьеры и оценивают его с более субъективной и негативной точки зрения.

Дискриминация в данном контексте может проявляться и в том, когда люди с одинаковыми характеристиками человеческого капитала или производительными способностями получают различную оплату в зависимости от их гендерного статуса или происхождения, местожительства и прочих других условий.

Каким образом предубеждения проникают в бесстрастный набор алгоритмов, которые работают с жесткими, холодными данными? ИИ хорош исключительно настолько, насколько «хороши» данные, которые его «питают». Его качество зависит от того, насколько хорошо создатели запрограммировали его думать, принимать решения, учиться и действовать. В результате ИИ может «унаследовать» или даже усилить предубеждения своих создателей, которые часто не знают о них, или ИИ может использовать предвзятые данные.

В этой связи возникает закономерный вопрос – кто же является создателем ИИ, а по сути – кто принимает решения? Возможно, мало кто замечает, но в РФ и в мире в целом зарождается новый цифровой класс по отношению к цифровым средствам производства. В этот класс входят те, кто имеет свободный доступ к персональным данным граждан, например работники многофункциональных центров или органы записи актов гражданского состояния, обладающие доступом к большим базам персональных данных граждан. К этому классу относятся также программисты, которые пишут программы для создания баз данных; системные администраторы, которые организуют их работу, ИТ-директора (руководители в сфере ИТ) и чиновники, которые всем этим управляют. Однако здесь и возникает главная сложность, потому что у чиновника, который осуществляет это управление, часто недостаточно развиты или полностью отсутствуют цифровые компетенции.

У программистов и системных администраторов есть некоторое чувство свободы. Так как они не давали присяги, однако, возможно, подписывали определенные обязательства о неразглашении, то ответственность несут только административную, а не уголовную. Чиновник свое управленческое решение принимает, основываясь на тех данных, которые ему предоставляет его представитель. Изменить, проверить эти данные или повлиять на них он не может в силу своей цифровой некомпетентности.

Обобщая портрет современного создателя алгоритмов для ИИ, мы можем отметить, что это люди в основном с техническим образованием, возраст которых колеблется от 20 до 30 лет; их считают технократами без особых убеждений. По словам И.С. Ашманова, новое поколение программистов принадлежит к так называемым «цифровым варварам», которые знают только цифровую сферу, не интересуясь при этом историей, культурой, этикой. Вся этика для них сосредоточена в алгоритмизации жизни[3].

Если у программистов слабо развиты или отсутствуют представления об этике, то распространение информации о другом человеке, с их точки зрения, не является кражей или преступлением. Системы ИИ принимают решения этического характера, потому что решения о людях – это этика, которая закладывается программистами, которые могут ею не обладать. Даже если авторы программ заявляют о том, что в основе их алгоритмов лежат нейтралитет и инклюзия, тем не менее, есть вероятность, что они разрабатывают их по чьему-то заказу, и в данном случае велика опасность, что эти нейтралитет и инклюзия продиктованы лицами, являющимися заказчиками самой программы. ИИ обладает способностью формировать решения отдельных лиц даже без их ведома, предоставляя тем, кто контролирует алгоритмические решения, полную неявную власть.

Кроме вопросов общекультурных знаний, создатели алгоритмов и сборщики данных, используемых для их тестирования и запуска, также не в состоянии предусмотреть все варианты развития того или иного события [7]. Приведем простой пример автомобиля без водителя, управляемого роботом. В этом случае, например, создатели робота могут забыть проверить распознавание им изображений в темное время суток в условиях сильного тумана в сельской местности.

Результаты применения технологий ИИ уже в нынешнем времени предоставляют возможность всем заинтересованным лицам (корпорациям-монополистам, правительству и др.) осуществлять сбор, хранение и анализ большого объема данных. Эти сведения совершенно беспрепятственно могут быть использованы с целью так называемого повышения эффективности деятельности и увеличения прибыли, при этом возможные последствия технологических прорывов и правительственных инноваций для

[3] Ашманов И.С. Чудовищная антиутопия в реальности. Режим доступа: http://новости-россии.ru-an.info/новости/началась-чудовищная-антиутопия-в-реальности-или-искусственный-интеллект-наступает/ (дата обращения: 21.01.2024).

отдельных групп населения останутся неучтенными (умышленно или неумышленно – неизвестно). При этом отдельный человек, живой, наделенный волей, эмоциями, желаниями и потребностями, не будет в курсе происходящего. Таким образом, делается шаг в сторону общества, где нет места индивидууму, где алгоритмы пишет сам ИИ и решения принимают роботы. Они, конечно, будут стремиться принимать решения, соответствующие предпочтениям большинства, но обратная сторона этих алгоритмизированных решений – невозможность выйти за определенные этим решением рамки. Это особенно опасно для молодого поколения, для которого приобретение опыта действовать самостоятельно, согласно собственному мнению будет практически недоступно.

Создатели алгоритмов для ИИ не могут охватить каждый элемент данных, который представляет обширность личности и потребности, желания, надежды этого человека. Неизвестно, кто сегодня собирает данные. Люди, которые отражают точки данных, могут не знать, для каких целей используются эти данные, или просто согласиться с условиями предоставления услуг, потому что у них не было реального выбора. Неизвестно также, кто зарабатывает деньги на этих данных и как они обрабатываются, для каких целей, чтобы их оправдать. На сегодняшний момент отсутствует прозрачность в таких вопросах, а надзор за использованием данных не регламентирован.

В будущем, основанном на алгоритмах, созданных ИИ, может произойти разрыв между людьми, разбирающимися в цифровых технологиях (которые являются наиболее востребованной в создаваемой цифровой экосистеме категорией), и теми, кто не обладает цифровой компетенцией или, в силу различных собственных причин, не хочет ею овладевать. Сами же алгоритмизированные решения смогут моментально спровоцировать разногласия любого рода между разными группами населения с помощью средств массовой информации, так как ИИ знает практически все о предпочтениях самих групп, классифицированных не только по какому-либо признаку, но и по предпочитаемому способу получения информации (телевидение, интернет, социальные сети и т.п.).

Традиционно дискриминация, вызываемая ИИ, связывается с угрозой массовой безработицы и всеми вытекающими из этого последствиями. Действительно, если рабочая задача может быть эффективно представлена алгоритмом, то ее легко выполнит машина.

ПОСЛЕДСТВИЯ РАСПРОСТРАНЕНИЯ ПРЕДВЗЯТОГО ИИ

Итак, сформулируем основные результаты, которые мы получили, определяя возможность дискриминации в новую цифровую эпоху. Нами определены и обозначены явные и латентные проблемы последствий распространения предвзятого ИИ. Явной, или основной, проблемой мы называем непосредственно дискриминацию человеческого капитала, вызванную некоторой некомпетентностью создателей той или иной технологии ИИ. Среди так называемых латентных, или сопутствующих, проблем мы выделяем следующие:

– алгоритмическая непрозрачность;
– уязвимость в области кибербезопасности вследствие отсутствия защиты от угроз новых мошенников в сети;
– несправедливость и предвзятость;
– отсутствие конкуренции;
– неблагоприятные последствия для работников;
– нарушение конфиденциальности и защиты данных и, как следствие, возможное нанесение вреда репутации человека, отсутствие ответственности разработчиков и пользователей за ущерб и отчетности в вопросах использования данных.

Далее нами выявлены возможные угрозы дискриминации в результате распространения ИИ и представлена их сущностная характеристика (см. таблицу).

Обществу, вероятно, потребуется дополнительное регулирование для защиты справедливости и прав человека в области ИИ. Однако регулирование ИИ в целом не является однозначно правильным подходом, поскольку использование систем ИИ слишком разнообразно для единого набора правил. Следует также учитывать национальные, отраслевые, географические и иные особенности при составлении такого рода правил. Необходимо проводить больше исследований, а также устраивать больше обсуждений и дискуссий на данную тему.

Таблица

Возможные угрозы дискриминации в результате распространения ИИ и проявление этих угроз

Возможная угроза дискриминации в результате распространения ИИ	**Проявление угрозы**
Преобладание данных, алгоритмов и прогнозного моделирования над человеческими суждениями и эмоциями	1) невозможность учета широких характеристик и особенностей каждой личности; 2) алгоритмы ИИ, разрабатываемые для компании, стремятся максимизировать прибыль, а не общественное благо; 3) у лиц, владеющих доступом к управлению ИИ и базами данных, появляются возможности для манипулирования людьми; 4) исчезновение личной конфиденциальности; 5) отсутствие контроля и прозрачности действий; 6) критика алгоритмов ИИ будет принижена, пресечена и отвергнута из-за преобладания цифровой логики над процессом; 7) люди теряют свою свободу воли в связи с необходимостью следовать алгоритму
Алгоритмически организованные системы ИИ содержат предвзятость	1) алгоритмы для ИИ разрабатываются с использованием данных, отобранных определенными привилегированными участниками в интересах таких же потребителей, как они сами; 2) программисты, создающие алгоритмы для ИИ, являются нерепрезентативной подгруппой населения; 3) ИИ ценит эффективность выше справедливости; 4) производители алгоритмов ИИ (корпорации и правительства) настраивают их таким образом, чтобы делать выбор, благоприятный для них самих
ИИ углубляет различия	1) пользователи, помещенные «на карантин» в различных идеологических областях, могут потерять способность человека к сопереживанию; 2) те, кто не является активным пользователем ИИ, окажутся в невыгодном положении; 3) негативные последствия будет иметь все, что алгоритмы считают рискованным или менее прибыльным; 4) массовый рост производительности за счет автоматизации увеличит неравенство между работниками и владельцами капитала
Рост безработицы в результате распространения ИИ	1) ИИ умнее, эффективнее, производительнее и дешевле, чем работник, которому необходимо создавать рабочие условия и гарантировать соблюдение его прав; 2) нарушение экономической модели рынка, согласно которой капитал обменивается на рабочую силу для обеспечения экономического роста (в случае, если рабочая сила перестанет быть частью этой модели)

Составлено авторами по материалам исследования

Еще одним результатом нашего исследования мы считаем разработку рекомендаций по минимизации и избеганию предвзятости и дискриминации в результате широкоформатной гражданской цифровизации и распространения ИИ.

Одной из первых причин, которая может породить предвзятость ИИ и усугубить различия между целыми группами людей, можно назвать отсутствие цифровой грамотности и цифровых компетенций у большого количества населения, не говоря уже об отсутствии знаний о действии механизма принятия решений ИИ и разработке алгоритмов на основе больших данных. Поэтому развивать цифровые компетенции необходимо начинать массово и с самого раннего возраста, вводя в обязательную программу государственного образования, чтобы выработать у широкой общественности понимание того, как функционируют алгоритмы ИИ.

Следующими мерами являются обеспечение прозрачности информации о том, как собираются и используются данные; развитие общественного понимания того, кто несет ответственность за их применение и нераспространение. Несмотря на огромный рост киберпреступности, практически неизвестны

факты уголовного преследования и наказания за них. По данным Банка России, за 2020 г. мошенничество «в цифровом объеме операций без согласия клиента» выросло на 38 %, а сумма украденных денег – на 52 % за год и составила 10 млрд руб.[4]

Люди сегодня проявляют весьма большой интерес, например, к тому, где и при каких условиях производятся продукты питания или одежда. Точно также следует задаваться вопросом о том, как собираются персональные данные, мнения в каких-либо опросах и, главное, кто принимает впоследствии решения. Необходимо знать, какова последовательность передачи этой информации, разрешаются ли допущения, какие критерии использовались для отбора сведений и данных и насколько они релевантны, а также какие стороны заинтересованы в принятии решений и насколько влиятельны эти стороны [8]. Другими словами, на данный момент лишь очень немногие понимают и осознают действие тех технологий ИИ, которые способны не только создавать, но и изменять существующую реальность. Однако, как мы уже отмечали, те, кто создает и развивает алгоритмы, не несут ответственности перед обществом. Необходимо в ближайшее время преодолеть это обстоятельство и разработать подход, который будет направлен на то, чтобы обязать разработчиков ИИ учитывать права человека на каждом этапе разработки. В свою очередь данный шаг выступит в качестве гарантии того, что алгоритмы, внедренные в обществе, будут устранять, а не усугублять социальное неравенство.

Механизмы надзора и контроля должны включать более строгие протоколы доступа к данным и обязательный перечень ответственных лиц с указанием их уровня ответственности и заключением договоров о неразглашении. Необходимо предусмотреть возможность дистанционного наблюдения за повторным обращением к использованным сведениям отдельным ответственным лицом, функции отказа системы, установление сроков доступа, невозможность продажи информации третьим лицам без согласия контролирующих органов. Кроме того, многие законодатели и контролирующие органы сегодня уже заявляют, что обширные серверные фермы социальных сетей должны стать более прозрачными и понятными. Эти монополисты имеют размеры, масштабы и в некотором смысле важность атомных электростанций и нефтеперерабатывающих заводов, но при этом почти не подвергаются надзору со стороны регулирующих органов. Данное положение должно измениться.

СПОСОБЫ ИЗБЕГАНИЯ ЦИФРОВОЙ ПРЕДВЗЯТОСТИ

Таким образом, можно обобщить изложенное выше и сформулировать следующее требование, позволяющее избежать цифровую предвзятость – алгоритмическая прозрачность должна быть установлена в качестве основополагающего требования для принятия всех решений на основе ИИ.

Отметим еще один нюанс. Алгоритмическая подотчетность, на наш взгляд, – это масштабный проект, требующий привлечения разнообразных специалистов и представителей общественности. Признание предвзятости часто является вопросом перспективы, и люди с разной расовой или любой иной идентичностью и экономическим происхождением заметят разные предубеждения. Создание разнообразных команд поможет снизить потенциальный риск проявления предвзятости ИИ. Команда создателей алгоритмов в идеале должна состоять из специалистов по обработке данных и бизнес-лидеров, представителей правительства, а также профессионалов с различным образованием и опытом, таких как юристы, бухгалтеры, социологи и специалисты по этике, журналисты, религиозные деятели. У каждого будет свое собственное представление об угрозе предвзятости и о том, как помочь ее смягчить.

Оценка прогнозируемых моделей на основе решений ИИ должна обязательно включать контроль и мониторинг со стороны социальных групп. Извлекая уроки из приведенных выше примеров, необходимо стремиться к тому, чтобы такие показатели, как истинная точность и частота ложноположительных результатов, были согласованными при сравнении различных социальных групп независимо от пола, этнической принадлежности или возраста.

ЗАКЛЮЧЕНИЕ

В заключение отметим, что и проведенное авторами исследование, и многочисленные публикации, и возникающие ситуации в сфере ИИ свидетельствуют о том, что возникла существенная необходимость последовательного регулирования вопросов применения ИИ на законодательном уровне.

[4] Банк России. Центробанк России. Обзор отчетности об инцидентах информационной безопасности при переводе денежных средств. Режим доступа: https://cbr.ru/statistics/ib/review_1q_2023/ (дата обращения: 25.01.2024).

Причем здесь снова нужно отметить указанные ранее требования: такие решения должны принимать лица, обладающие высоким уровнем развития цифровой грамотности, целесообразно включать в команды разработчиков представителей самых разных профессий, а процессу принятия решений надлежит происходить на принципах открытости, доступности и прозрачности. Руководителям на самом высоком уровне нужно обладать пониманием необходимости ответственного ИИ, то есть этичного, надежного, безопасного, хорошо управляемого, совместимого и объяснимого. Это требование выдвигается и в выступлении Президента РФ В.В. Путина[5]. Дискриминация человеческого капитала, вызванная действием ИИ, не является неизбежной, так как все зависит от человека и от того, как нация, цивилизованное общество смогут положить ей конец.

Список литературы / References

1. *Leonov V.A., Kashtanova E.V., Lobacheva A.S.* Ethical aspects of artificial intelligence use in social spheres and management environment. European Proceedings of Social and Behavioural Sciences. 2021;118:989–998. http://doi.org/10.15405/epsbs.2021.04.02.118
2. *Suen H.-Y., Hung K.E., Lin Ch.-L.* Intelligent video interview agent used to predict communication skill and perceived personality traits. Human-centric Computing and Information Sciences. 2020;3(10). http://dx.doi.org/10.1186/s13673-020-0208-3
3. *Vinichenko M.V., Narrainen G.S., Melnichuk A.V., Chalid Ph.* The influence of artificial intelligence on human activities. In: Frontier information technology and systems research in cooperative economics. Heidelberg: Springer; 2020. Pp. 561–570.
4. *Symitsi E., Stamolampros P., Daskalakis G., Korfiatis N.* The informational value of employee online reviews. European Journal of Operational Research. 2021;2(288):605–619. http://dx.doi.org/10.1016/j.ejor.2020.06.001
5. *Sinha N., Singh P., Gupta M., Singh P.* Robotics at workplace: an integrated Twitter analytics – SEM based approach for behavioral intention to accept. International Journal of Information Management. 2020;55:102210. http://dx.doi.org/10.1016/j.ijinfomgt.2020.102210
6. *Courtland R.* Bias detectives: the researchers striving to make algorithms fair. Nature. 2018;558:357–360. https://doi.org/10.1038/d41586-018-05469-3
7. *Popkova E.G., Gulzat K.* Contradiction of the digital economy: public well-being vs. cyber threats. In: Economy: complexity and variety vs. rationality: Proceedings of the 9th National Scientific and Practical Conference, Vladimir, April 17–18, 2019. Cham: Springer; 2019. Pp. 112–124.
8. *Heinrichs B.* Discrimination in the age of artificial intelligence. AI & Society. 2022;37:143–154. https://link.springer.com/article/10.1007/s00146-021-01192-2

[5] Дружинин А. Путин: лидер в сфере искусственного интеллекта станет властелином мира. Режим доступа: https://ria.ru/20170901/1501566046.html (дата обращения: 26.01.2024).