Информация для цитирования:

Харитонов Ю. С., Савина В. С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник Пермского университета. Юридические науки. 2021. Вып. 53. С. 488–515. DOI: 10.17072/1995-4190-2021-53-488-515.

Kharitonova Yu. S., Savina V. S., Pagnini F. *Predvzyatost' algoritmov iskusstvennogo intellekta: voprosy etiki i prava* [Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues]. *Vestnik Permskogo universiteta. Juridicheskie nauki* – Perm University Herald. Juridical Sciences. 2021. Issue 53. Pp. 488–515. (In Russ.). DOI: 10.17072/1995-4190-2021-53-488-515.

УДК 347.2

DOI: 10.17072/1995-4190-2021-53-488-515

**ПРЕДВЗЯТОСТЬ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:
ВОПРОСЫ ЭТИКИ И ПРАВА**

Исследование выполнено в рамках Программы развития Междисциплинарной научно-образовательной школы Московского университета «Математические методы анализа сложных систем»

Ю. С. Харитонов

Московский государственный университет имени М. В. Ломоносова

E-mail: sovet2009@rambler.ru

В. С. Савина

Российский экономический университет имени Г. В. Плеханова

E-mail: savin-viktoriya@yandex.ru

Ф. Паньини

LOYTEC Electronics GmbH (Вена, Австрия)

E-mail: fabric3p@gmail.com

Поступила в редакцию 03.03.2021

Введение: в работе рассматриваются правовые проблемы применения технологии искусственного интеллекта для решения социально-экономических задач. Конвергенция двух подрывных технологий – искусственного интеллекта (Artificial Intelligence) и Data Science (работа с большими данными (англ. Big Data) – породила качественное изменение общественных отношений в различных сферах жизни людей. Преобразующую роль сыграли как классические области искусственного интеллекта: алгоритмическая логика, планирование, представление знаний, моделирование, автономные системы, мульти-агентные системы, экспертные системы, системы поддержки принятия решений, моделирование, распознавание образов, обработка изображений и обработка естественного языка, так и специальные области: обучение представлениям, машинное обучение, оптимизация, статистическое моделирование, математическое моделирование, аналитика данных, обнаружение знаний, наука о сложности, вычислительный интеллект, анализ событий и поведения, анализ социальных медиа / сетей, а в последнее время – глубокое обучение и когнитивные вычисления. Названные технологии искусственного ин-

теллекта и больших данных применяются в различных сферах бизнеса для упрощения и ускорения принятия разного рода и значимости решений. В то же время установлено, что самообучающиеся алгоритмы создают или воспроизводят неравноправие участников оборота, приводят к дискриминации в связи с обнаруживающейся алгоритмической предвзятостью. **Цель:** сформировать представление о направлениях правового регулирования алгоритмической предвзятости при применении технологии искусственного интеллекта с позиций права на основе анализа российских и зарубежных научных концепций. **Методы:** эмпирические методы сравнения, описания, интерпретации; теоретические методы формальной и диалектической логики; частнонаучные методы: юридико-догматический и метод толкования правовых норм. **Результаты:** искусственный интеллект имеет множество преимуществ (улучшение творческих способностей, услуг, безопасности, образа жизни, помощь в решении проблем), но в то же время вызывает и множество опасений в связи с неблагоприятным воздействием на автономию личности, конфиденциальность, а также основные права и свободы человека. Алгоритмическая предвзятость существует даже тогда, когда у разработчика алгоритма нет намерения дискриминации, и даже когда рекомендательная система не принимает на вход демографическую информацию: тем не менее, тщательно используя сходство товаров и пользователей, алгоритм может в конечном счете рекомендовать товар очень однородному составу пользователей. Представляется, что выявленные проблемы и риски предвзятости при применении технологии искусственного интеллекта должны быть учтены юристами и разработчиками и максимально смягчены на уровне как формирования этических принципов и требований, так и правовой политики и права на национальном и наднациональном уровнях. Юридическое сообщество полагается на решение проблемы алгоритмической предвзятости через принятие разного рода деклараций, политик и стандартов, соблюдение которых будет необходимо при разработке, тестировании и эксплуатации систем искусственного интеллекта. **Выводы:** если оставить без внимания предвзятые алгоритмы, то они приведут к решениям, могущим повлечь коллективное разрозненное влияние на определенные группы людей даже без намерения программиста проводить различие. Современное изучение предполагаемых и непредвиденных последствий применения алгоритмов искусственного интеллекта исключительно важно, особенно потому, что меры текущей государственной политики могут быть недостаточными для выявления, смягчения и устранения последствий такой неочевидной предвзятости для участников правовых отношений. Решение проблем предвзятости алгоритмов посредством только технических средств не приведет к желаемым результатам. Мировое сообщество задумалось о введении стандартизации и разработке этических принципов, что позволит установить рамки для справедливого применения искусственного интеллекта в принятии решений. Необходимо создание специальных норм, устанавливающих ограничения алгоритмической предвзятости. Независимо от сфер выявления таких нарушений они должны содержать общие признаки недобросовестного поведения участников общественных отношений, которое может быть квалифицировано как нарушение прав человека или нарушение добросовестной конкуренции. Минимизации предвзятости алгоритмов будет способствовать обязательное введение в оборот данных таким образом, что не позволит проводить явную или скрытую сегрегацию различных групп населения, т. е. анализу подлежат должны данные без явных групповых признаков в полном их разнообразии. Общеприменимая модель искусственного интеллекта тем самым будет строиться на анализе данных всех социально-правовых групп общества.

Ключевые слова: искусственный интеллект; предвзятость алгоритмов; права человека; самообучающееся программное обеспечение; взаимодействие человека с машиной; риски применения искусственного интеллекта; стандартизация систем искусственного интеллекта

ARTIFICIAL INTELLIGENCE'S ALGORITHMIC BIAS: ETHICAL AND LEGAL ISSUES

The research was prepared within the framework of the Program for Development of the Interdisciplinary Scientific and Educational School of the Moscow University "Mathematical Methods of Analysis of Complex Systems"

Yu. S. Kharitonova

Lomonosov Moscow State University

E-mail: sovet2009@rambler.ru

V. S. Savina

Plekhanov Russian University of Economics

E-mail: savin-viktoriya@yandex.ru

F. Pagnini

LOYTEC Electronics GmbH

E-mail: fabric3p@gmail.com

Received 03.03.2021

Introduction: *this paper focuses on the legal problems of applying the artificial intelligence technology when solving socio-economic problems. The convergence of two disruptive technologies – Artificial Intelligence (AI) and Data Science – has created a fundamental transformation of social relations in various spheres of human life. A transformational role was played by classical areas of artificial intelligence such as algorithmic logic, planning, knowledge representation, modeling, autonomous systems, multiagent systems, expert systems (ES), decision support systems (DSS), simulation, pattern recognition, image processing, and natural language processing (NLP), as well as by special areas such as representation learning, machine learning, optimization, statistical modeling, mathematical modeling, data analytics, knowledge discovery, complexity science, computational intelligence, event analysis, behavior analysis, social network analysis, and also deep learning and cognitive computing. The mentioned AI and Big Data technologies are used in various business spheres to simplify and accelerate decision-making of different kinds and significance. At the same time, self-learning algorithms create or reproduce inequalities between participants in circulation, lead to discrimination of all kinds due to algorithmic bias. **Purpose:** to define the areas and directions of legal regulation of algorithmic bias in the application of artificial intelligence from the legal perspective, based on the analysis of Russian and foreign scientific concepts. **Methods:** empirical methods of comparison, description, interpretation; theoretical methods of formal and dialectical logic; special scientific methods such as the legal-dogmatic method and the method of interpretation of legal norms. **Results:** artificial intelligence has many advantages (it allows us to improve creativity, services and lifestyle, to enhance the security, helps in solving various problems), but at the same time it causes numerous concerns due to the harmful effects on individual autonomy, privacy, and fundamental human rights and freedoms. Algorithmic bias exists even when the algorithm developer has no intention to discriminate, and even when the recommendation system does not accept demographic information as input: even in the absence of this information, due to thorough analysis of the similarities between products and users, the algorithm may recommend a product to a very homogeneous set of users. The identified problems and risks of AI bias should be taken into consideration by lawyers and developers and should be mitigated to the fullest extent possible, both when developing ethical principles and requirements and in the field of legal policy and law at the national and supranational levels. The legal community sees the opportunity to solve the problem of algorithmic bias through various kinds of declarations, policies, and standards to be followed in the development, testing, and operation of AI systems. **Conclusions:** if left unaddressed, biased algorithms could lead to decisions*

that would have a disparate collective impact on specific groups of people even without the programmer's intent to make a distinction. The study of the anticipated and unintended consequences of applying AI algorithms is especially necessary today because the current public policy may be insufficient to identify, mitigate, and remedy the effects of such non-obvious bias on participants in legal relations. Solving the issues of algorithmic bias by technical means alone will not lead to the desired results. The world community recognizes the need to introduce standardization and develop ethical principles, which would ensure proper decision-making with the application of artificial intelligence. It is necessary to create special rules that would restrict algorithmic bias. Regardless of the areas where such violations are revealed, they have standard features of unfair behavior of the participants in social relations and can be qualified as violations of human rights or fair competition. Minimization of algorithmic bias is possible through the obligatory introduction into circulation of data in the form that would not allow explicit or implicit segregation of various groups of society, i.e. it should become possible to analyze only data without any explicit attributes of groups, data in their full diversity. As a result, the AI model would be built on the analysis of data from all socio-legal groups of society.

Keywords: artificial intelligence; algorithmic bias; human rights; robotic performance; self-learning software; human-machine interaction; risks of artificial intelligence applications; standardization of artificial intelligence systems

Введение

Наша жизнь количественно, а иногда и качественно оценивается скрытыми от глаз алгоритмами.

Конвергенция двух подрывных технологий – искусственного интеллекта (Artificial Intelligence) и Data Science (работа с большими данными (англ. Big Data) – породила качественное изменение общественных отношений в различных сферах жизни людей. Преобразующую роль сыграли как классические области искусственного интеллекта: алгоритмическая логика, планирование, представление знаний, моделирование, автономные системы, мультиагентные системы, экспертные системы, системы поддержки принятия решений, моделирование, распознавание образов, обработка изображений и обработка естественного языка, так и специальные области: обучение представлениям, машинное обучение, оптимизация, статистическое моделирование, математическое моделирование, аналитика данных, обнаружение знаний, наука о сложности, вычислительный интеллект, анализ событий и поведения, анализ социальных медиа / сетей, а в последнее время – глубокое обучение и когнитивные вычисления. Названные технологии искусственного интеллекта и больших данных применяются в различных сферах бизнеса для упрощения и ускорения принятия разного рода и значимости решений.

В статье исследуются различные способы, благодаря которым алгоритмы способствуют принятию необъективных решений, и оцениваются такие действия с позиции права и закона.

Алгоритмы используются для профилирования и прогнозирования поведения людей, предположительно нейтральным образом. Однако алгоритмы могут быть «испорчены» человеческими предубеждениями скрытым образом, значительно влияя на жизнь людей [34], приводя в конце концов к нарушению их конституционных прав и свобод, допуская откровенную дискриминацию.

Технология искусственного интеллекта для принятия решений: как работают алгоритмы?

Согласно некоторым определениям, искусственный интеллект (ИИ) – это семейство методов, посредством которых алгоритмы выявляют или изучают положения предсказательной силы, основанные на анализе больших данных¹. По существу, алгоритм – это пошаговая процедура решения проблемы.

Наиболее значимой формой искусственного интеллекта является машинное обучение,

¹ *Предсказательная сила* (англ. *predictive power*), согласно Википедии, – это термин философии науки, которым обозначается способность теории формулировать положения относительно еще не исследованных объектов и явлений физического мира исходя только из данных самой теории и без обращения к эмпирическим свидетельствам.

которое включает в себя методы глубокого обучения, основанные на нескольких уровнях представления данных и способные отображать сложные отношения между входящими и исходящими данными. Однако людям трудно интерпретировать алгоритмически «выученные» представления [43]. Под обучением представлению специалисты понимают изучение представлений входных данных, как правило, путем их преобразования или извлечения из них функций (некоторыми способами), что упрощает выполнение такой задачи, как классификация или прогнозирование [13].

Напомним, что Интернет является крупнейшим публичным хранилищем больших данных, созданных человечеством. Проблема заключается в том, что для принятия тех или иных решений использование данных из этого источника должно сопровождаться тщательной проверкой их достоверности. Иными словами, как говорят специалисты, качество данных имеет решающее значение. В свою очередь, одной из важных проблем качества данных является систематическое их искажение, проявляющееся в различных формах. Эти искажения влияют на алгоритмы машинного обучения, которые изначально разрабатываются для улучшения взаимодействия с пользователем. А в дальнейшем эта проблема еще больше усугубляется возникающей откровенной предвзятостью, которую добавляют эти алгоритмы в результате обработки входных данных к информации, лежащей в основу принятия решения. И все это возникает не только на фоне отсека, например, каких-либо кандидатов при принятии решений о трудоустройстве, но и в позитивном контексте, когда разного рода интернет-сервисами применяются системы рекомендаций и персонализации [9].

Следует отметить, что концепция «машинного обучения» не является абсолютно новой, фактически она восходит к 1950-м годам¹. Изначально сама идея машинного обучения возникла в академической, а не прикладной сфере, и сегодня фактически представляет собой совокупность идей и методов, пришедших из разных областей, таких как статистика, приклад-

ная математика, искусственный интеллект и др.² Если говорить более обобщенно, возможности глубокого обучения с помощью представлений основываются на следующих логических приемах. Как лучше всего объяснить ребенку, что такое бабочка или белка? Лучший способ – просто показать ему это, возможно, на картинке или на видео. Как лучше всего научить ребенка кататься на велосипеде или чистить зубы? Ему просто нужно приобрести опыт. Любое другое описание, каким бы точным оно ни было, не будет столь эффективным. Термин «машинное обучение» означает машины (в широком смысле – включая программы, алгоритмы), которые учатся на данных³. При этом существует множество различных типов данных: от физических величин, таких как параметры атмосферы, до данных о предпочтениях или вкусах людей, цифровых изображениях с камер, видео и т. д.

Вначале машинное обучение сильно отличалось от того, каким мы знаем его сегодня⁴: в первой версии, по сути, необходимо было предоставить компьютерам необходимые правила для анализа данных. Однако сколько этих правил, а точнее, сколько исключений пришлось бы учесть, чтобы позволить, например, алгоритму переводить с английского языка на китайский? Отдельно даже велась дискуссия относительно того, чтобы перевести в код все законы и подзаконные правовые акты, судебную и иную практику и решения уполномоченных органов [12] для того, чтобы создать комфортную среду реализации смарт-контрактов с учетом частноправовых и публично-правовых правил торговли, оказания услуг, защиты прав потребителей, соблюдения требований сертификации продукции и налогообложения проведенных операции и т. д. Но, вероятно, исключений было бы так много, что практически невозможно было бы построить что-то работающее правильно и при этом действительно полезное: именно поэтому успех

² *Brief Timeline of AI*. URL: <https://www.livescience.com/47544-history-of-a-i-artificial-intelligence-infographic.html>.

³ *Types of Machine Learning Algorithms You Should Know*. URL: <https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861>.

⁴ *Dartmouth Summer Research Project on Artificial Intelligence*. URL: <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1904/1802>.

¹ *Complete Historical Overview*. URL: <http://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf>.

того, что сейчас является де-факто отраслью компьютерной инженерии, не заставил себя ждать. Необходимо, чтобы «машина» сама могла создавать свои правила, подобно тому, как мы, люди, учимся и получаем опыт¹. Машинное обучение в значительной степени получило толчок к развитию с появлением технологий сбора, обработки и хранения больших данных (Big Data) и Интернета вещей (IoT, Internet of Things), использование которых предполагает налаживание обмен и сбор данных миллионами устройств. Только подумайте, какими данными мы обмениваемся с помощью наших смартфонов, постов в Facebook, Twitter, LinkedIn, видео на YouTube, телевизоров Smart TV и многого другого².

Точно так же у нас теперь есть процессоры, способные собирать и обрабатывать эти большие объемы данных за короткий промежуток времени: данные должны быть отобраны, распознаны, а затем проанализированы с помощью статистических и прогностических моделей, чтобы извлечь и выбрать полезную информацию, например, в маркетинговых целях, чтобы повлиять на продажи или определить возможные новые продукты для рынка и даже целые новые рынки³. Это примеры того, как можно использовать машинное обучение и искусственный интеллект для улучшения процесса принятия решений в бизнесе: начиная с продуктов, определенных машиной, компании могут фактически выбирать, что предложить своему потребителю на рынке. В мире, все больше управляемом данными, в мире, который бежит все быстрее и быстрее, необходимо иметь системы искусственного интеллекта, которые помогут нам принимать лучшие решения за меньшее время.

Но как это происходит, как машина может обучаться? По сути, существует три метода: контролируемое обучение (Supervised Learning), неконтролируемое обучение (Unsuper-

vised Learning) и обучение с подкреплением (Reinforcement learning)⁴. При использовании метода контролируемого обучения машина способна строить правила на основе большого количества предоставленных ей данных: в этом случае алгоритму нужны и данные для анализа, и полученные результаты, т. е. практически миллиарды примеров, на которых машина может учиться. В примере с освоением китайского языка потребуется очень большое количество английских предложений и их соответствующих китайских переводов, чтобы машина сама могла построить алгоритм перевода: чем больше у нее данных, тем большему она может научиться, построить модели, и чем точнее модели, тем точнее будут результаты.

Также можно взять пример из другой области применения: спам-фильтр электронной почты, по сути, является программой машинного обучения, которая может научиться отмечать спам после получения примеров спама и примеров обычных неспамовых писем: ценное подспорье для аналитиков кибербезопасности.

При «обучении без наблюдения» у нас просто есть большой набор данных, например все продажи Amazon в Европе, и мы можем проанализировать их, чтобы найти полезную информацию. Алгоритм будет искать корреляции, повторяющиеся закономерности между данными и затем возвращать полезную информацию: большинство людей, покупающих определенный товар, например смартфон, покупают и чехол. Подобные алгоритмы можно также встретить в крупных торговых сетях или в международных аэропортах: информация о покупках в сочетании с другими данными позволяет экстраполировать полезные закономерности. Полученные знания могут быть использованы в коммерческой деятельности, чтобы указать, как расположить товары на полках магазинов, чтобы привлечь внимание покупателей и тем самым побудить их к покупке, создать таргетированную рекламу.

В «обучении с подкреплением» идея заключается в том, что алгоритм учится на своих собственных ошибках: в этом случае мы не даем алгоритму больших объемов данных, на которых он должен учиться или в которых он

¹ *The Impact of AI on Cybersecurity*. URL: <https://www.computer.org/publications/tech-news/trends/the-impact-of-ai-on-cybersecurity>.

² *Can Machine Think? Richard Feynman on Artificial General Intelligence*. URL: <https://www.cantorsparadise.com/richard-feynman-on-artificial-general-intelligence-2c1b9d8aae31>.

³ *Applying Machine Learning to the Internet of Things*. URL: <https://medium.com/iotforall/applying-machine-learning-to-the-internet-of-things-5bd0216d4cc3>.

⁴ *An in-depth guide to supervised machine learning classification*. URL: <https://builtin.com/data-science/supervised-machine-learning-classification>.

должен искать закономерности, а оставляем его свободным совершать ошибки. Обучение с подкреплением – это еще один метод, который используем и мы, люди: когда ребенок учится ездить на велосипеде, он может потерять равновесие, часто падает, а затем снова встает. Это обучение, основанное на опыте. Метод «обучение с подкреплением» часто применяется сегодня для создания различных игр, начиная с классической игры в шахматы и заканчивая видеоиграми.

Для каждой из этих областей существует множество различных методик. Одна из широко применяемых основана на нейронных сетях. Вычислительные системы называются нейронными сетями только потому, что они вдохновлены тем, как работает наш мозг: биологические нейронные сети, состоящие из миллиардов взаимосвязанных нейронов. Немного похожие на русскую матрешку, нейронные сети являются частью машинного обучения, а машинное обучение – это подсистема искусственного интеллекта: фундаментальной частью интеллекта является способность учиться и, таким образом, совершенствоваться и расти.

Однако человеческий мозг – это очень сложная «машина», работающая на нескольких концептуальных уровнях, поэтому также было введено понятие глубокого обучения: набор техник и алгоритмов, основанных на нейронных сетях, организованных в слои или уровни. Таким образом, искусственный интеллект – это способность машины имитировать поведение человека, изучать новые концепции, моделировать поведение и принимать решения: интеллект означает способность адаптироваться. Искусственный интеллект имеет множество преимуществ (например, улучшение творческих способностей, услуг, безопасности, образа жизни, помощь в решении проблем), но в то же время вызывает множество опасений (неблагоприятное воздействие на автономию личности, конфиденциальность, а также основные права и свободы человека). При всех впечатляющих перспективах упрощения жизни общества, в связи с распространением решений на основе применения технологии искусственного интеллекта, исследователи провели большую работу по выявлению последствий применения данной

технологии [47]. Были не только выявлены достоинства технологии искусственного интеллекта, но и определены несколько категорий рисков самого разного характера¹:

- потеря работы из-за автоматизации;
- нарушение конфиденциальности;
- ‘Deepfakes’;
- алгоритмическая ошибка, вызванная неверными данными;
- социально-экономическое неравенство;
- автоматизация оружия.

Как можно видеть, обозначенные категории рисков пока не систематизированы в достаточной мере и скорее фиксируют наиболее острые проблемы социальной жизни. В то же время, на наш взгляд, наименее изученной и обсуждаемой в России и мире остается проблема социально-экономического неравенства, порождаемого применением технологии искусственного интеллекта.

Авторы настоящей работы сфокусировались на правовых проблемах применения технологии искусственного интеллекта для решения социально-экономических задач. В последнее время в литературе все отчетливее звучат указания на несправедливость [54], предвзятость [20] и дискриминацию со стороны искусственного интеллекта [28].

Представляется, что выявленные проблемы и риски применения технологии искусственного интеллекта должны быть учтены юристами и разработчиками и максимально смягчены как на уровне формирования этических принципов и требований, так и правовой политики и права на национальном и наднациональном уровнях.

Алгоритмическая предвзятость искусственного интеллекта в контексте дискриминации отдельных групп граждан

В 2016 году информационная интернет-платформа журналистских расследований ProPublica сообщила², что инструменты таргетированной рекламы Facebook позволяли рекламодателям исключать определенные группы,

¹ *Dangers of AI: 6 Risks Of Artificial Intelligence/ Mike Thomas/ January 14, 2019| Updated: April 1, 2021 | Built In: URL: <https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence>.*

² *Facebook Lets Advertisers Exclude Users by Race. URL: <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.*

которые были обозначены как носители «этнического родства» (например, всех, кто «близок» к афроамериканцам или американцам азиатского происхождения), из числа лиц, получающих услуги и информацию о рекламируемых продуктах и сервисах. Журналисты смогли разместить объявление о сдаче жилья в аренду с исключением из числа потенциальных арендодателей представителей различных групп и меньшинств (афроамериканцы, люди, интересующиеся пандусами для инвалидных колясок, и евреи). Это вызвало дискуссию о том, что алгоритмы Facebook потенциально нарушают Закон о справедливом жилищном обеспечении¹. В феврале 2017 г. Facebook опубликовал обновления своей рекламной политики и отчитался о создании инструментов для усиления запрета на дискриминацию в рекламе жилья, найма труда или кредитования. Однако повторное тестирование системы независимыми экспертами эту информацию не подтвердило. По-прежнему можно было размещать рекламу, нацеленную на участников Facebook, которые искали дом, и исключаящую всех, кто имел «сходство» с афроамериканцами, американцами азиатского происхождения или латиноамериканцами².

В данном случае мы имеем дело с такими параметрами алгоритма, которые были преднамеренно введены интересантами при размещении объявлений и не блокировались системой Facebook.

Однако, когда мы говорим о предвзятости алгоритма искусственного интеллекта, мы имеем в виду то, что не было прямо заложено в программу ее разработчиком, но возникло в силу самообучения алгоритма. Что в этом контексте понимать под предвзятостью?

Алгоритмическая предвзятость была обнаружена в 1970–80-е годы в практике Медицинской школы больницы Св. Георгия в Соединенном Королевстве, которая использовала компьютерную программу для первоначального отбора абитуриентов. Программа, которая

имитировала выбор, сделанный приемной комиссией в прошлом, отказала в собеседовании шестидесяти кандидатам, потому что они были женщинами или имели неевропейские имена. Важно, что программный код не был работой какого-то программиста-злоумышленника. Напротив, предвзятость возникла в процессе приема кандидатов потому, что была фактически заложена в данные ранее принятых решений. Компьютерная программа по сути лишь усугубила проблему и придала ей видимость объективности. Комиссия Великобритании по расовому равенству признала Медицинскую школу Св. Георгия виновной в расовой и половой дискриминации при приеме абитуриентов в 1988 году [26].

Можно полагать, что это был один из первых выявленных случаев, показавший, что некритичный импорт логических алгоритмов оптимизации процесса управления и принятия решений из сферы техники и инженерии в некоторые области социальных наук порождает *алгоритмическую предвзятость* в значительных масштабах. Однако за прошедшее время данная проблема только усугубилась, но не была разрешена на законодательном уровне.

В 2015 году сервис Google Photo случайно пометил фотографии двух афроамериканцев как горилл [11]. Спустя три года после этого шокирующего инцидента Google, пытаясь исправить ситуацию, подверг цензуре слова «горилла», «шимпанзе», «обезьяна» по поисковым запросам и тегам изображений [52]. В свое оправдание компания ссылалась на несовершенство технологии распознавания фотографий, маркировки изображений, семантического значения слов.

Позже расследование алгоритмической несправедливости³ было проведено группой ученых, которые измеряли точность трех коммерческих алгоритмов гендерной классификации на своей платформе Parliaments Benchmark, который был сфокусирован на распознавании лиц по полу и типу кожи. Исследователи аннотировали набор данных с помощью системы классификации кожи Фитцпатрика [24] и протестировали эффективность гендерной классификации на четырех подгруппах: смуглые женщи-

¹ Закон о справедливых жилищных условиях запрещает дискриминацию арендодателей и покупателей жилья со стороны домовладельцев, продавцов и кредиторов по признаку их расы, цвета кожи, религии, сексуальной ориентации, национальности, инвалидности или семейного положения // Fair Housing Act. URL: <https://www.justice.gov/crt/fair-housing-act-1>.

² URL: <https://www.documentcloud.org/documents/3191165-Facebook-Propublica-Ad.html>.

³ Investigating Algorithmic Injustice.

ны, смуглые мужчины, светлые женщины и светлые мужчины. В результате все классификаторы показали наилучшие результаты для более светлых людей и мужчин в целом. Классификаторы показали худшие результаты для темных женщин. Эксперимент с применением технологии искусственного интеллекта показал, что сохраняется существенный разрыв в восприятии сервисами компьютерного зрения индивидуумов в зависимости от пола и типа кожи [15].

Однако представляется, что проблема заключается не только в несовершенстве технологии распознавания картинки, но и в том, что всем этим технологиям и до сих пор в значительной степени не достает способности понимать контекст или абстрактные концепции в процесс принятия решений. И даже самым лучшим алгоритмам не хватает способности использовать здравый смысл или абстрактные концепции, чтобы уточнить свою интерпретацию мира, как это делают люди. Примечательно, что ситуация вызвала обеспокоенность: Google необходимо быть более прозрачным в своих методах устранения алгоритмической предвзятости, а также в отношении технической незрелости существующих систем распознавания объектов.

В 2018 году нашумевшим делом в отношении нарушения трудовых прав и дискриминации женщин стала история с Amazon. Специалисты по машинному обучению компании Amazon.com Inc AMZN.О обнаружили серьезную проблему: их новый механизм рекрутинга «не любил» женщин. Самообучаемая система Amazon проанализировала введенные данные о принятых ранее на различные позиции кандидатах и «научилась» тому, что кандидаты-мужчины предпочтительнее. Amazon была вынуждена отменить применение алгоритмов искусственного интеллекта для подбора персонала, который продемонстрировал предвзятое отношение к женщинам. В то же время этот критерий отбора не был намеренно встроен в систему; он стал результатом машинного обучения. Такое видение программы возникло именно в результате статистики по найму мужчин и женщин, которой компания располагала на момент обучения. В целом, примеры, демонстрирующие дискриминацию раз-

личных групп граждан, говорят о том, что искусственный интеллект, не обладающий предвзятками сам по себе, будет принимать решения в рамках существующего в обществе статус-кво, т. е. фиксировать то, что происходит в мире людей¹. Как утверждает в своем исследовании Т. Шаулова, «по некоторым оценкам, в отборе и найме персонала станет доминировать меритократический подход, когда параметры “пола” и “возраста” перестанут определять выбор работодателя. В связи с этим можно предположить, что гендерная предвзятость при найме и продвижении работников исчезнет, а обсуждение вопроса равенства полов будет выглядеть диссонансом. Но на практике конкуренция за рабочие места сохраняет гендерный характер: при найме на работу большинство работодателей продолжают руководствоваться гендерными стереотипами. Ожидаемой “трансформации менталитета” не произошло. Остается надежда на новые технологии» [51]. Такие надежды на цифровые технологии возлагаются в связи с беспристрастностью и объективностью алгоритмов [1; 22; 39; 16; 8; 31] в отличие от предвзятости и ограниченности экспертного мнения². Однако приведенные примеры говорят об обратном. Трудно оспорить утверждение о том, что «предвзятость “реального мира” формирует и предвзятость алгоритмов» [51].

Злоупотребления при применении технологий машинного обучения и искусственного интеллекта проявляются и в иных контекстах: в медицине – через призму справедливости и защиты жизни пациентов [6; 7], в сфере научных исследований – применительно к фейковым, ничем не подтвержденным результатам исследований [42], и др.

Так, в литературе отмечается, что системы здравоохранения столкнутся с тремя пробле-

¹ URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.

² Women matter. URL: <https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence>; Обзор рынка труда и заработных плат в России в 2018 году (Hays Salary Guide 2018). URL: <https://hays.ru/>; Запрет на указание требований к полу и возрасту в вакансиях. URL: <https://www.superjob.ru/research/about/>; Работодатели стали реже обращать внимание на семейное положение соискателей, но гендерные предпочтения остались. URL: <https://www.superjob.ru/research/about/>.

мами при устранении алгоритмической предвзятости. Во-первых, отсутствие четких определений и стандарта «справедливости»; во-вторых, недостаточная контекстная специфичность, в-третьих, «черный ящик» природы алгоритмов [44]. Здесь вопрос справедливости и предвзятости поднимается по сути на новый этический уровень. Ученые подчеркивают, что подход заботы о справедливости особенно актуален, когда алгоритмы используются для поддержки полярных решений (т. е. когда один полюс прогноза приводит к решениям, которые обычно более желательны, чем другие). Например, когда прогнозы используются для распределения ограниченных ресурсов здравоохранения, в группе пациентов, которой это могло бы принести пользу. Возникают вопросы определения критериев справедливости, в вопросах здравоохранения часто приобретающих взаимную несовместимость. Даже когда модели используются для уравнивания пользы и вреда для принятия оптимальных решений для отдельных лиц (для неполярных решений), а соображения справедливости неуместны, модели, данные или проблемы выборки могут привести к предвзятым прогнозам, поддерживающими решения, которые дифференцированы по принципу «вредно/полезно» для разных групп [46].

Обсуждаемый в литературе случай применения алгоритмов для прогнозирования объема необходимой медицинской помощи [56] показывает, что не всегда происходит дискриминация только по расовому или половому признаку. Так, несмотря на единую методику расчета для всех пациентов, алгоритм считал чернокожего пациента менее нуждающимся в медицинской помощи, чем белого, даже если у первого больше объективных причин получить медпомощь. Но в коде не было заложено преимущество белокожих пациентов, и алгоритм работал правильно. Ошибочной была исходная идея разработчиков, что равные расходы на медпомощь свидетельствуют об одинаковой потребности в ней, поэтому алгоритм рассчитывал рекомендации на основании расходов пациентов на медпомощь в прошлом. Однако расходы человека на медицинские услуги сильно зависят от уровня дохода и социального положения. Следовательно, алго-

ритм закрепил дискриминацию, существовавшую еще до него: пациенты, которые в прошлом получали меньше медпомощи из-за низкого уровня дохода, окажутся обделены ею и в будущем.

Также были выявлены не очевидные на первый взгляд нарушения непредвзятости и справедливости и в сфере научных исследований. Вопросы возникли применительно к фейковым, ничем не подтвержденным результатам исследований [42], которые приводили к увеличению рейтингов ученых, журналов, исследовательских организаций и строились по принципам, сходным с принципами отбора кандидатов на работу или абитуриентов. Подрыв в сфере научных исследований сопровождался генерацией вымышленных авторов, включением ученых, которые не имели никакого отношения к статье в качестве соавтора, и даже «заваливанием» журналов материалами из некачественного мусора, написанного искусственным интеллектом.

Важно, что в обозначенной выше подборке примеров принимаемое алгоритмом искусственного интеллекта решение несло в себе и порождало дальнейшую предвзятость в широком смысле как относящуюся к результатам, которые систематически менее благоприятны для отдельных лиц в определенной группе и где нет существенных различий между группами, оправдывающих такой вред¹.

Если оставить без внимания предвзятые алгоритмы, они могут привести к решениям, которые могут иметь коллективное разрозненное влияние на определенные группы людей даже без намерения программиста проводить и внедрять такое разделение. Своевременное изучение предполагаемых и непредвиденных последствий применения алгоритмов искусственного интеллекта необходимо особенно потому, что меры текущей государственной политики могут быть недостаточными для выявления, смягчения и устранения последствий такой неочевидной предвзятости для участников правовых отношений.

¹ *Understanding Bias in Algorithmic Design* // Impact.Engineered, September 6, 2017. URL: <https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e>.

Алгоритмическая предвзятость и рекомендательные системы: скрытая цена рыночной власти¹

В настоящее время одним из наиболее инновационных для ведения предпринимательской деятельности подходов является дата – управляемый процесс принятия решений (*data-driven decision making*) [40]. Данная технология осуществляется в рамках применения алгоритмов искусственного интеллекта и означает, что субъекты всегда опираются на данные, чтобы принимать решения и что каждый элемент бизнес-стратегии основан на интерпретации данных искусственным интеллектом [55]. Подход, основанный на данных, позволяет компаниям использовать свои массивы данных для получения уникальной информации о клиентах.

Индустрия, основанная на данных [36], создала рекомендательные системы, упрощающие и ускоряющие принятие решений. Это технологии, которые на основе информации о профиле пользователя предсказывают, какие объекты (фильмы, музыка, книги, новости, веб-сайты) будут ему интересны. Рекомендательные системы широко используются в сфере развлечений, цифровом маркетинге, биомедицинской индустрии и др.

Ранее нами было установлено, что рекомендательная система на основе заранее собранных и обработанных данных описывает возможные пользовательские предпочтения. Выводы алгоритма служат важной основой для дальнейших рекомендаций пользователю того или иного контента. Этот эффект используется как часть механизма принятия решений в бизнесе и государственном управлении, а также весьма востребован в экономике как современный инструмент маркетинга (продвижения) для прогнозирования продаж, внедрения продуктов и др. Поскольку некоторые потребители и компании нуждаются в рекомендации или прогнозе для будущего бюджета, координации труда и цепочки поставок, динамические системы рекомендаций крайне востребованы сегодня для точного прогнозирования [5].

Таким образом, происходит взаимное влияние информационных потоков, «погло-

щаемых» пользователем Интернет-сети, и особенностей этого «поглощения», под влиянием которых на абстрактном уровне формируются не личные поведенческие закономерности, а уже привычки всего социума, определяющие, в свою очередь, правила поведения отдельного человека в социуме. Как было показано, алгоритмическая предвзятость [30] выявляется даже тогда, когда у разработчика алгоритма нет намерения дискриминации и даже когда рекомендательная система не принимает на вход демографическую информацию: тем не менее, тщательно используя сходство товаров и пользователей, алгоритм может в конечном итоге рекомендовать товар очень однородной группе пользователей [23].

Когда искусственный интеллект рекомендует, какой новый телесериал будет для нас наиболее интересен, или сообщает банку, стоит ли предоставить ссуду определенному человеку, в рекомендации не указываются причины, по которым предлагается такой результат. Алгоритм просто опирается на предыдущие закономерности. Определенному проценту людей, которые смотрели многие из сериалов, входящих в базу данных, понравился этот новый сериал, так что пользователю, вероятно, сериал тоже понравится. Или: более 70 % людей определенного возраста, гражданского статуса, средней заработной платы и проживающих в районах похожи на лицо, подающее заявку на ссуду, в конечном итоге не вернет ее, поэтому лучше не предоставлять ему денежные средства в долг.

Напомним, предвзятость может быть не только естественной, образовавшейся случайно из-за особенностей входящих данных, но и искусственной, заложенной намеренно, например в виде предпочтения интересов неких третьих лиц. Примером подобной предвзятости может служить небольшое намеренное изменение маршрута пользователя на карте в навигационной системе, чтобы он проехал или прошел мимо определенной точки, например мимо магазина, заказавшего рекламу².

Следует отметить, что поток информации, новостей, рекламы в социальных медиа и на развлекательных платформах организован не сбалансированным образом, а с помощью algo-

¹ *Inequality: A hidden cost of market power*. URL: <https://www.oecd.org/competition/inequality-a-hidden-cost-of-market-power.htm>.

² URL: https://ethics.cdto.center/3_3.

ритмов, которые созданы не только в интересах пользователей, но во многом для максимального использования самой платформы. Алгоритмическая предвзятость в коммерческой сфере проявляется в том смысле, что искусственно создается фрагментация мнений, чем и усиливается поляризация мнений и оценок продуктов и сервисов. Это артефакт онлайн-платформ, также называемый «алгоритмической сегрегацией» [33; 32].

Особенностью предвзятости применительно к функционированию цифровых платформ является следующее. Поток информации, поступающей к нам через социальные медиа и онлайн-платформы, оптимизируется не содержанием или актуальностью информации, а популярностью и близостью к цели. Обычно это выполняется для максимального использования платформы. В качестве побочного эффекта это приводит к алгоритмической предвзятости, которая, как считается, усиливает фрагментацию и поляризацию общественных дебатов [53], но может играть роль и при обсуждении коммерческих продуктов и сервисов.

Важными кажутся результаты исследования, посвященного продвижению новостей в социальных медиа. Представляется, что некоторые выводы могут быть экстраполированы и на коммерческие сервисы. Так, поток новостей в новых медиа выбирается не по ценности информации, а по популярности, по «лайкам» [45]. Поскольку люди склонны отождествлять себя со взглядами, подобными их собственным, и с большей вероятностью будут положительно отмечать, проставляя «лайк», соответствующие новости, поставщики услуг заинтересованы в том, чтобы уже целенаправленно направлять информацию [4]. Это означает, что пользователи даже не сталкиваются с информацией о товарах и сервисах, отличных от их любимых. При этом нарастает разработка таких вот «эффективных» алгоритмов, чтобы обеспечить поток с наибольшим шансом собрать максимальное количество реакций, «лайков», в конечном счете – покупок.

Другой важный фактор, действующий в том же направлении, связан с функцией «поделиться», которая в значительной степени отвечает за быстрое распространение новостей и иной информации и, таким образом, за повышение популярности товаров, работ, услуг. Так,

в новостной сфере это распространение происходит в социальной сети, где ссылки формируются в основном в результате гомофилии, т. е. обмен информацией происходит между людьми с похожими взглядами. Разнообразие взаимодействий человека (семья, школа, работа, хобби и т. д.) может способствовать диверсификации источников информации [10], хотя, например, в отношении политических взглядов гомофилия проявляется особенно сильно [19]. В контексте распространения дезинформации в Facebook и Твиттере было показано, что распространение информации среди пользователей с похожими убеждениями приводит к эффекту эхо-камеры [21; 27]. Потребление новостей в Facebook порождает очень резко очерченные медиа-кластеры, которые пользователи не покидают [48]. Потребляя эти новости, пользователи становятся все более поляризованными, особенно вокруг определенных СМИ и стран [49]. Тем не менее в литературе можно встретить указание на то, что «связь между фрагментацией и поляризацией мнений, а также алгоритмической предвзятостью онлайн-платформ на сегодняшний день не доказана, а выводы различных исследований противоречивы» [41].

Рекомендации и персонализация – полезные технологии, которые все больше и больше влияют на наши повседневные решения. Однако предвзятость, которая существует в реальном мире и отражается в обучающих данных, может быть смоделирована и усилена рекомендательными системами и в конечном счете возвращена в виде предвзятых рекомендаций пользователям. Возникает так называемая «самовоспроизводящаяся петля обратной связи ИИ» (*self-perpetuating social feedback loops in AI-based predictive policing*) [17], которая постепенно усиливает значение внешних обстоятельств – фильтров, в которых мы живем. Необъективные рекомендации могут также усиливать стереотипы, например, основанные на гендерной или этнической принадлежности, что может привести к неравному воздействию. Применительно к рекомендательным системам исследователями была выявлена связь между предсказуемостью чувствительных характеристик и предвзятостью результатов рекомендаций, что позволило определить предел предвзятости рекомендаций, основанный на этой связи,

и даже разработать FaiRecSys – алгоритм, который смягчает алгоритмическую предвзятость путем постобработки матрицы рекомендаций с минимальным влиянием на полезность рекомендаций, предоставляемых конечным пользователям [28].

Это свидетельствует о том, что разработчики программного обеспечения пытаются формализовать ограничение справедливости и цены товара при выдаче рекомендаций, чтобы добиться выработки наиболее «справедливых» рекомендаций. Применяются специальные методы устранения статистической погрешности или моделирования такой погрешности с целью повышения эффективности работы рекомендательных систем и выявления тенденций (например, в отношении того, чему пользователи дают более высокие оценки или как пользователи, которые изменяют свои базовые оценки с течением времени, оказывают влияние на производительность рекомендательных систем) [38]. Некоторые из этих статистических смещений (например, временное) являются частью динамики интереса к рекомендуемым товарам и должны быть смоделированы соответствующим образом. Существуют также наборы инструментов с открытым исходным кодом, например AI Fairness 360 Open Source Toolkit¹, который помогает пользователям изучать, сообщать и смягчать дискриминацию и предвзятость в моделях машинного обучения на протяжении всего жизненного цикла приложения, основанного на использовании технологии искусственного интеллекта. Он использует семьдесят «показателей справедливости» и десятки современных алгоритмов снижения предвзятости, разработанных исследовательским сообществом. То есть в целом разработчики и заказчики технологических решений стараются принимать некоторые меры для смягчения неизбежно возникающей алгоритмической предвзятости.

Однако решение проблем предвзятости алгоритмов исключительно техническими средствами не приведет к желаемым результатам. Мировое сообщество задумалось о введении стандартизации и разработке этических принципов, позволяющих установить рамки для справедливого применения искусственного интеллекта в принятии решений.

Этика и искусственный интеллект: правовые подходы в разных государствах

Проблема предвзятости алгоритмов отчасти решается с помощью регуляторных инструментов. В первую очередь во многих странах были приняты этические принципы и стандарты, которые адресованы технологическим решениям в сфере искусственного интеллекта.

Этика помогает принимать правильные решения в том или ином контексте на основе имеющейся информации. Таким образом, необходимо анализировать этические ценности и принципы, думая о том, как действовать, чтобы их применить. Например, в программном документе Агентства ЕС по основным правам (FRA) описывается потенциал дискриминации отдельных лиц с помощью алгоритмов и говорится, что «принцип недискриминации, закрепленный в статье 21 Хартии основных прав Европейского Союза, необходимо учитывать при применении алгоритмов в повседневной жизни»².

Современная эволюция систем искусственного интеллекта привела к обширному обсуждению этики искусственного интеллекта. Т. Хагендорф [29] провел анализ и сравнение двадцати двух руководств, нормативных принципов и рекомендаций, направленных на то, чтобы избавить системы, построенные на применении технологии искусственного интеллекта, от недостатков. Ранее А. Джобин с соавторами [35] проанализировали, возникает ли глобальное согласие по основным этическим вопросам на основе некоторых руководящих принципов искусственного интеллекта. Их исследование выявило глобальное сближение вокруг пяти этических принципов, а именно: 1) прозрачность; 2) справедливость и честность; 3) нематериальность; 4) ответственность и 5) конфиденциальность. Вместе этические принципы и стандарты позволяют увидеть достаточно четкую картину возможного бедственного положения, в котором сегодня оказались методы искусственного интеллекта с позиции человеческих ценностей, поскольку эти позиции говорят о необходимости исправить технологии до установленного этического уровня.

² EU Agency for Fundamental Rights (FRA) (2018) #BigData: Discrimination in data-supported decision making. URL: <http://fra.europa.eu/en/publication/2018/big-data-discrimination>.

¹ URL: <https://aif360.mybluemix.net/>.

Поиск надлежащих инструментов регламентации процессов функционирования систем искусственного интеллекта осуществляют не только ученые в сфере технологии, но и законодатели, как на национальном, так и на международном уровнях. В период с 2017 по 2018 год было опубликовано более семидесяти рекомендаций по этике искусственного интеллекта [58].

В первую очередь во многих странах были приняты этические принципы и стандарты, которые адресованы технологическим решениям в сфере искусственного интеллекта. Европейским парламентом в 2017 г. был опубликован Отчет Европейского парламента с рекомендациями для Комиссии по гражданскому праву правила робототехники, в котором, в частности, было отмечено, что «тенденция к автоматизации требует, чтобы те, кто занимается разработкой и коммерциализацией приложений искусственного интеллекта, изначально строили систему безопасности и этики, тем самым признавая, что они должны быть готовы принять юридическую ответственность за качество производимой ими технологии» (пункт N)¹. Также в данном документе подчеркивается важность сбалансированного подхода, предполагающего необходимость стандартизации аппаратного и программного обеспечения систем искусственного интеллекта, и в то же время недопустимость создания препятствий инновациям.

Основными принципами Стратегии единого цифрового рынка², утвержденными Европейской комиссией, являются:

1) повышение доступа отдельных потребителей, а также организаций к онлайн-товарам и услугам (что предполагает ликвидацию разрыва между онлайн и офлайн-средой, с тем чтобы устранить барьеры для трансграничной торговли);

2) создание надлежащих условий для развития цифровых сетей и услуг (для этого необходимы высокоскоростные безопасные инфраструктуры и контент-сервисы, а также адекватное нормативно-правовое регулирование инноваций, инвестиций и обеспечение добросовестной конкуренции);

3) максимальный рост потенциала цифровой экономики (это предполагает инвестиции в инфраструктуры и информационно-компьютерные технологии, включая облачные вычисления и большие данные, а также повышение качества государственных услуг).

Как отмечено в Рекомендации Парламентской ассамблеи Совета Европы № 2102 от 28 апреля 2017 г. «Слияние с технологиями, искусственный интеллект и права человека»³, законодателям все труднее адаптироваться к темпу развития науки и технологий и разрабатывать необходимые нормативные акты и стандарты (пункт 3), в связи с чем следует разработать руководящие принципы, в частности, по вопросу автоматической обработки операций, направленных на сбор, обработку и использование персональных данных, а также общие рамки стандартов (пункт 9.1). В данном документе сделан важный вывод о том, что любая машина, любой робот или любой искусственный интеллект должны оставаться под контролем человека.

В рамках Евразийского экономического союза принят технический регламент «О безопасности машин и оборудования»⁴, разрабатываются схемы, удостоверяющие, что системы алгоритмических решений не демонстрируют неоправданной предвзятости. Например, сейчас развивается и совершенствуется Стандарт рассмотрения алгоритмических предубеждений (IEEE p7003)⁵ – это один из стандартов этики IEEE (разрабатывается в рамках Глобальной инициативы IEEE по этике автономных и интеллектуальных систем). Он направлен на пре-

¹ Report of the European Parliament with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL)) dated as of 27th of January, 2017. URL: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN>.

² Political Guidelines for the next European Commission – A New Start for Europe: My Agenda for Jobs, Growth, Fairness and Democratic Change (15 July 2014). URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192>.

³ Recommendation № 2102 (2017) of Parliamentary Assembly of the Council of Europe «Technological convergence, artificial intelligence and human rights», 28 April 2017. URL: <http://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=23726&lang=en>.

⁴ Решение Комиссии Таможенного союза от 18 окт. 2011 г. № 823 (ред. от 16.05.2016) [Электронный ресурс]. Доступ из справ.-правовой системы «КонсультантПлюс».

⁵ P7003 – Algorithmic Bias Considerations. URL: <https://standards.ieee.org/project/7003.html>.

доставление отдельным лицам или организациям, создающим алгоритмические системы, среды разработки, правил, как избежать непреднамеренных, необоснованных и неуместно различающихся результатов для пользователей.

Элементы Стандарта включают: набор рекомендаций по тому, что нужно делать при разработке или использовании алгоритмических систем, следуя принципиальной методологии взаимодействия с заинтересованными сторонами (людьми); определение и обоснование целей использования алгоритма, принципов, которые фактически заложены в алгоритмическую систему (продукт); практическое руководство для разработчиков, позволяющее определить, когда им следует отступить, чтобы оценить возможные проблемы предвзятости в их продукте, и указание на методы, которые они могут использовать для решения задачи минимизации предвзятости; процедуры сравнительного анализа и критерии для выбора валидационных наборов данных для контроля качества смещения; методы установления и передачи информации и границы применения, для которых система была разработана, и валидации, для защиты от возникающих непредвиденных нежелательных последствий, возникающих в результате применения алгоритмов за пределами указанных границ; методы управления ожиданиями пользователей для уменьшения погрешности из-за неправильной интерпретации результатов работы системы пользователями; и другие, которые еще предстоит определить [37]. То есть по существу зачастую создатели стандартов и правил создают руководства для разработчиков, закладывая в них этические и нормативные ограничения.

В течение последнего десятилетия основные положения вышеобозначенных и иных стратегических, программных документов были внедрены в национальное законодательства целого ряда стран: в частности, Великобритания, Германия, Китай, США, Япония, Южная Корея предпринимают значительные усилия в отношении регламентации базовых стандартов в сфере робототехники и искусственного интеллекта¹.

¹ *Draft Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*, 31.05.2016 / Committee on Legal Affairs / European Parliament; Rapporteur: Mady Delvaux. URL: <http://www.europarl.europa>

В целом ряде государств были приняты национальные планы (стратегии) по развитию цифровых технологий. Например, в США был принят Национальный стратегический план исследований и разработок в области искусственного интеллекта², в КНР – национальный стратегический план и промышленная политика Коммунистической партии Китая «Сделано в Китае 2025 г.: глобальные амбиции, основанные на локальной защите»³, в Японии – Новая стратегия роботизации⁴ и т. д.

Японский Комитет по политике в сфере робототехники в качестве меры, позволяющей решить либо предотвратить проблемы, связанные с функционированием роботов нового поколения, назвал в том числе разработку стандартов проектирования и производства роботов, предполагающих определение ответственности производителей за вред, причиненный действиями роботов⁵. Также были приняты «Этические руководящие принципы для Японского общества искусственного интеллекта (JSAI)» (2017)⁶, где отмечена необходимость искусственного интеллекта соблюдать этические правила.

Российская Стратегия развития информационного общества в Российской Федерации, утвержденная Президентом РФ в 2017 г.⁷, определяет искусственный интеллект как комплекс технологических решений. Следовательно, данные решения могут быть объектами стандартизации, поскольку, в соответствии

[eu/sides/getDoc.do?pubRef=-//ep/nongsm1%2bcomparl%2bpe-582.443%2b01%2bdoc%2bpdf%2bv0/en](http://eu.sides/getDoc.do?pubRef=-//ep/nongsm1%2bcomparl%2bpe-582.443%2b01%2bdoc%2bpdf%2bv0/en). P. 4.

² *The National Artificial Intelligence Research and Development Strategic Plan*. URL: https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.

³ *Made in China 2025: global Ambitions Built on Local Protections*. URL: https://www.uschamber.com/sites/default/files/final_made_in_china_2025_report_full.pdf.

⁴ *New Robot Strategy. Japan's Robot Strategy – Vision, Strategy, Action Plan*. URL: http://www.meti.go.jp/english/press/2015/pdf/0123_01b.pdf.

⁵ *Japan Ministry of Economy, Trade and Industry Robot Policy Council. Robot Policy Council Report, May 2006 (in Japanese)*. Цит. по: *Weng Y.H., Chen C.H., Sun C.T.* The legal crisis of next generation robots: on safety intelligence // *Proceedings of the 11th international conference on Artificial intelligence and law*. 2007. P. 206.

⁶ URL: http://ai-elsi.org/wp-content/uploads/2017/02/pdf_англ.пер.TheJapaneseSocietyforArtificialIntelligenceEthicalGuidelines. URL: <http://ai-elsi.org/wp-content/uploads/2017/05/JSAIEthical-Guidelines-1.pdf>.

⁷ *О Стратегии развития информационного общества в Российской Федерации на 2017–2030 годы: указ Президента РФ от 9 мая 2017 г. № 203 // Собр. законодательства Рос. Федерации. 2017. № 20, ст. 2901.*

с подпунктом 6 статьи 2 Федерального закона от 29 июня 2015 г. № 162-ФЗ (ред. от 03.07.2016) «О стандартизации в Российской Федерации», такими объектами могут являться продукция (работы, услуги), процессы, системы менеджмента, терминология, условные обозначения, исследования (испытания) и измерения (включая отбор образцов) и методы испытаний, маркировка, процедуры оценки соответствия и иные объекты. Данный подход соответствует и учитывает основные положения международных документов.

Методические рекомендации по внедрению интернета вещей для оптимизации контрольно-надзорной деятельности¹ содержат определение понятия «промышленный интернет вещей, применяемый в контрольно-надзорной деятельности»: это совокупность автоматических или автоматизированных средств измерения, передачи и обработки данных, систем реагирования и дистанционного мониторинга, обеспечивающих контрольно-надзорные органы достоверными сведениями о состоянии проверяемых объектов. Достоверность этих сведений может быть обеспечена только с помощью стандартизации. В данных рекомендациях также содержатся предложения об отказе от избыточного сбора персональных данных, ограничения его сообразно целям обработки, а также от установления регламентированного порядка их раскрытия, допустимых сценариев агрегации и деперсонификации.

В 2020 году Федеральным агентством по техническому регулированию и метрологии была разработана Перспективная программа стандартизации по приоритетному направлению «Искусственный интеллект» на период 2021-2024 годы², предусматривающая разработку 217 стандартов в данной сфере.

В соответствии с данной Программой в 2020 году был принят, в частности, ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения», в котором: определено понятие

доверия к системам искусственного интеллекта; приведена классификация факторов, влияющих на качество и способность систем искусственного интеллекта вызывать доверие на стадиях жизненного цикла; формализована взаимосвязь качества и способности систем искусственного интеллекта вызывать доверие; приведена классификация основных способов обеспечения доверия к системам искусственного интеллекта³. Вместе с тем данный стандарт не может быть использован для систем «сильного» или «общего» искусственного интеллекта.

В 2021 году в Российской Федерации был принят Национальный стандарт системы искусственного интеллекта. Классификация систем искусственного интеллекта (ГОСТ Р 59277-2020б, принят 3 января 2021)⁴. Целью данного документа стало установление принципов классификации систем искусственного интеллекта. Внедрение этого стандарта необходимо для повышения эффективности использования систем искусственного интеллекта при решении прикладных задач. Установление классификации систем искусственного интеллекта позволит сравнивать различные решения по таким параметрам, как: тип деятельности, структура знаний, функции контура управления, безопасность, конфиденциальность, степень автоматизации, методы обработки информации, интеграция / взаимодействие, сложность системы, архитектура, специализация. В том числе данный стандарт содержит классификацию в зависимости от соответствия следующим классам конфиденциальности: (0) Открытая информация: (1) Внутренняя информация: (2) Конфиденциальная информация: (3) Секретная информация.

В то же время отмечено, что классы можно характеризовать исходя из различных дополнительных аспектов или подклассов, например:

- наличие/отсутствие внешнего наблюдения, осуществляемого человеком-оператором либо другой автоматизированной системой;
- степень понимания системы;
- степень реактивности / отзывчивости;
- уровень устойчивости функционирования;

¹ Методические рекомендации по внедрению и использованию промышленного интернета вещей для оптимизации контрольной (надзорной) деятельности: утв. протоколом заседания проектного комитета от 9 нояб. 2017 г. № 73(13) [Электронный ресурс]. Доступ из справ.-правовой системы «КонсультантПлюс».

² URL: <https://www.economy.gov.ru/material/file/28a4b183b4aee34051e85ddb3da87625/20201222.pdf>.

³ URL: https://www.securitymedia.ru/pic/%D0%A1%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D1%8B_%D0%98%D0%98.pdf

⁴ URL: https://allgosts.ru/35/020/gost_r_59277-2020.

- степень надежности и безопасности;
- вид аппаратной реализации;
- степень приспособляемости к внутренним или внешним изменениям;
- способность оценивать свою собственную работоспособность / пригодность;
- способность принимать решения и планировать.

На международном уровне в настоящее время разработаны следующие стандарты: ISO / МЭК 20546: 2019 Информационные технологии. Большие данные; ISO / IEC TR 20547-2: 2018 Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования; ISO / МЭК 20547-3: 2020 Информационные технологии. Эталонная архитектура больших данных. Часть 3. Эталонная архитектура; ISO / IEC TR 20547-5: 2018 Информационные технологии. Эталонная архитектура больших данных. Часть 5. Дорожная карта стандартов.

Вазль Уильям Диаб, председатель SC 42 (Искусственный интеллект), на семинаре данного комитета по искусственному интеллекту и общедоступным данным отметил: «Растущий спрос на системы искусственного интеллекта для понимания бизнес-проблем способствует росту прогнозов, таких как прогнозы IDC, расходы на когнитивную деятельность и искусственный интеллект вырастут до 52,2 млрд долларов в 2021 году, достижение среднегодового темпа роста (CAGR) – в 46,2 % за прогнозируемый период 2016–2021 гг. Экосистема созрела для стандартизации» [57]. В данном докладе виден широкий подход к стандартизации, который выходит за рамки традиционных проблем совместимости, отмечается, что такие соображения, как стоимость, устойчивость, безопасность и конфиденциальность, играют наиболее важную роль при определении требований к системам искусственного интеллекта. Некорректная работа систем искусственного интеллекта сопряжена с риском для жизни и здоровья людей, значительным экономическим и экологическим ущербом.

Наиболее важной для предотвращения и преодоления предвзятости алгоритмов искусственного интеллекта представляется нормативная регламентация этического принципа справедливости (честности) применительно к функ-

ционированию систем искусственного интеллекта.

Не случайно в упоминавшемся выше Отчете Европейского парламента с рекомендациями для Комиссии по гражданскому праву правила робототехники¹ подчеркивается необходимость обеспечения недискриминации, надлежащей правовой процедуры, прозрачности и понятности в процессах принятия решений. Правила, регулирующие, в частности, ответственность, прозрачность и подотчетность, признаны безусловно полезными, поскольку они отражают, по сути, европейские и универсальные гуманистические ценности, характеризуют вклад Европы в развитие общества. В то же время отмечено, что они не должны влиять на процесс исследований, инноваций и разработок в области робототехники. В Отчете указывается, что руководящие этические рамки функционирования роботов должны основываться на принципах благотворительности, непричинения вреда, автономии и справедливости, на принципах и ценностях, закрепленных в статье 2 Договора о Европейском союзе² и в Хартии основных прав, таких как человеческое достоинство, равенство, справедливость, недискриминация, информированное согласие, личная и семейная жизнь и защита данных, а также другие основополагающие принципы и ценности законодательства Союза: отсутствие стигматизации, прозрачность, автономия, индивидуальность ответственности и социальной ответственности, а также существующих этических норм и кодексов. При этом особое внимание следует уделять роботам, которые представляют значительную угрозу конфиденциальности из-за их размещения в традиционно защищенных и частных сферах, а также из-за того, что они могут извлекать и отправлять личные и конфиденциальные данные.

В российском стандарте ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положе-

¹ Report of the European Parliament with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL)) dated as of 27th of January, 2017. URL: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN>.

² Договор о Европейском союзе (Маастрихт, 07.02.1992) (в ред. Лиссабонского договора 2007 г.). URL: <https://eu-law.ru/treaties/teu/>.

ния»¹ раскрываются последствия несоблюдения этического принципа справедливости, такие как предвзятость и необъективность (*bias*). Под ними понимается свойство системы искусственного интеллекта, заключающееся в принятии ошибочных решений, связанных со статистической смещенностью обучающей выборки исходных данных.

Факторы снижения качества могут быть связаны с естественными (непреднамеренное снижение качества) или искусственными (преднамеренное снижение качества) причинами. Примерами преднамеренного снижения качества, специфичными для систем искусственного интеллекта, являются:

1) на стадии создания системы – наличие преднамеренных искажений в обучающей выборке системы распознавания изображений, приводящих к ошибкам в работе системы распознавания, вызванным специальными, заранее определенными искажениями в исходных данных, включая «состязательные» атаки;

2) на стадии эксплуатации системы – отсутствие достоверных и представительных оценок устойчивости системы распознавания изображений к воздействию преднамеренных «состязательных» атак, приводящее к неустойчивой работе системы в процессе ее эксплуатации.

Примерами непреднамеренного снижения качества систем искусственного интеллекта являются:

1) на стадии создания системы – использование статистически смещенной обучающей выборки, приводящей к появлению «предвзятостей» в результатах работы системы;

2) на стадии эксплуатации системы – нарушение конфиденциальности обрабатываемых данных в условиях, когда уровень конфиденциальности данных существенно и неконтролируемо возрос в процессе эксплуатации системы искусственного интеллекта вследствие накопления и обобщения информации.

Для предотвращения предвзятости и обеспечения доверия к системам искусственного интеллекта на соответствующих стадиях жизненного цикла разработчиками стандарта предлагаются, например, следующие решения:

– выбор представительного набора существенных характеристик системы и корректных правил их определения;

– формирование представительной обучающей выборки (например, для биометрических систем искусственного интеллекта) (ГОСТ Р 57194.1);

– очистка набора данных различными способами;

– статистический анализ наборов исходных данных и оценка их представительности и качества;

– кросс-валидация выборки, полученной при разметке данных людьми;

– непосредственная корректировка модели;

– наложение ограничений на допустимую область применения системы искусственного интеллекта.

Данные способы открывают достаточно полный доступ к процессу создания системы (основная часть требований разработчика должна быть открытой, т. е. подлежащей проверке любой заинтересованной стороной), что может быть затруднительно для разработчика системы. Кроме того, перечисленные способы показательны для разработчика, но, как правило, неинформативны для потребителя системы.

Цифровая трансформация отраслей изменила ландшафт стандартизации, ключевыми вопросами стали не только технические требования, в том числе обеспечение надежности систем искусственного интеллекта, но также этические нормы, социальный кластер, потребности бизнеса, политики и т. д. Все эти разноплановые вопросы должны быть трансформированы в технические требования, а это возможно лишь посредством создания основополагающих стандартов, которые позволяют опираться на единообразную терминологию, использовать унифицированные руководства по применению, а также эталонные меры.

В перспективе в процессе стандартизации необходимы: более детальная разработка терминологии в сфере искусственного интеллекта, стандартов для систем искусственного интеллекта с использованием машинного обучения, управления рисками, нормативно-технического регулирования предвзятости в системах искусственного интеллекта и принятия решений с его помощью, надежности искусственного интеллекта; стандарт оценки надежности нейрон-

¹ URL: https://www.securitymedia.ru/pic/%D0%A1%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D1%8B_%D0%98%D0%98.pdf.

ных сетей; обзор примеров использования, этических и социальных проблем, вычислительных подходов для систем искусственного интеллекта; структура управления процессами для анализа больших данных; анализ управленческих последствий использования искусственного интеллекта организациями.

При использовании систем искусственного интеллекта целесообразно учитывать, что необходимо решать вопросы публично-правового характера: получения согласия на обработку данных от субъектов этих данных; определения правосубъектности этих лиц; установления юридической ответственности в связи с недобросовестным использованием полученных данных для принятия решений. Стандартизация может помочь в их решении. Стандарты помогут смягчить проблемы этики искусственного интеллекта и снизить градус вполне понятной озабоченности общества по поводу предвзятости данных систем, обеспечивая их эффективное и быстрое внедрение.

На наш взгляд, при разработке, тестировании и эксплуатации систем искусственного интеллекта необходима стандартизация, разработка норм технических стандартов. Заметим, что при формулировании этических принципов применительно к системам искусственного интеллекта следует избегать ассоциирования данных систем с субъектами права, наделяя их определенным объемом правосубъектности. Мы придерживаемся подхода, изложенного в Стратегии 2030 и исходим из того, что искусственный интеллект – это технология, а не субъект права. Система регуляторов должна быть гибкой, она не должна ограничивать развитие технологий искусственного интеллекта и не препятствовать их эффективному функционированию, а также развитию этических принципов в данной сфере.

**Применимые правовые решения
для смягчения последствий
применения технологии
искусственного интеллекта**

Предвзятость систем искусственного интеллекта может быть ограничена стандартами и этическими принципами, закрепленными в стратегиях и иных программных документах. Однако огромное количество разного рода установленных стандартов и этических правил

вызвало скепсис в отношении практического эффекта от их внедрения. Л. Флориди [25] выделил пять неэтических рисков, связанных с увеличением количества рекомендаций, касающихся этики, в приложениях искусственного интеллекта:

1. Этический шопинг: это практика, используемая для оправдания себя в отношении решений, выбора и процессов, принятых компаниями. В качестве «меню» принципов и ценностей выбираются наиболее подходящие для конкретных целей компании. Таким образом, вместо того, чтобы компании были нацелены на поиск в направлении изменения в поведении, изменения и исправления алгоритмов и изменений в обработке данных, они ориентированы на выбор принципов, которые легитимизируют их поведение.

2. Этический Bluewashing (декларация приверженности компании социально значимым целям): речь идет о принятии косметических мер, чтобы создать видимость существования этических принципов у технологических компаний. Акцент делается на распространение рекламы и маркетинг, чтобы создать видимость, что компания занимается этическими вопросами. Однако фактически ничего не делается, поскольку реализация эффективных мер по исправлению процессов или решений может повлечь расход многих ресурсов;

3. Этическое лоббирование: относится к использованию саморегулирования через этические кодексы, чтобы задержать или предотвратить разработку соответствующего законодательства, обосновать ограничение в части его соблюдения и ослабить правоприменение. Компании не видят преимуществ этичного поведения в отношении своей политики лоббирования [14] и предпочитают «избегать хорошего и необходимого законодательства (или его исполнения) о проектировании, разработке и внедрении цифровых процессов, продуктов, услуг или других решений» [25].

4. Этический демпинг: означает экспорт неэтичных исследований в страны, где нет сильного законодательства, запрещающего проведение подобных практик и импорт результатов этих исследований в страну происхождения компании, которая их проводила. Термин «этический демпинг» был введен Единством науки с обществом и для общества

Европейской комиссии в 2013 году. Наконец, существуют две причины экспорта неэтичной практики: намеренная эксплуатация из стран с высоким уровнем дохода в страны с низким или средним уровнем дохода и эксплуатация из-за отсутствия знаний и этических проблем [50, pp. 1–4].

5. Этическое «отлынивание» (уклонение от соблюдения этических норм): это проведение небольшого количества мероприятий, направленных на применение этических принципов и ценностей, когда отдача от них невелика, а стоимость ответственности перекладывается на третьих лиц. Термин «отлынивание» пришел с финансового рынка и означает «тенденцию делать меньше работы, когда прибыль меньше»¹.

Представляется, что проблема правоприменения заключается в том, что в некоторых случаях правила могут быть даже закреплены в законе, но их толкование и применение подвергаются серьезному переосмыслению в связи с развитием и внедрением технологии искусственного интеллекта.

Поэтому важно все-таки не ограничиваться утверждением только стандартов и принципов. Отсутствие правовой определенности в вопросе искусственного интеллекта создает возможность того, что предвзятые алгоритмы могут считаться юридически допустимыми, в то время как подходы, разработанные для исправления предвзятости, могут считаться незаконно дискриминационными. Например, недавно предложенное правило Министерства жилищного строительства и городского развития Техаса («HUD») (США)² установило бы первый пример нормативного определения США для алгоритмической дискриминации, создало бы безопасную гавань от ответственности за последствия дискриминации в сфере жилья.

Федеральное правительство США предоставляет налоговые льготы застройщикам, которые строят так называемое «жилье с низким доходом». Эти кредиты администрируются уполномоченными государственными органами, и предпочтение отдается развитию жилищ-

ного строительства в районах, где проживает население с низким доходом. The Inclusive Communities Project – это некоммерческая организация из Техаса, помогающая малообеспеченным семьям получить доступное жилье. В 2008 году они подали иск против техасского агентства, ответственного за администрирование этих налоговых льгот, утверждая, что оно несоразмерно распределяло слишком много налоговых льгот «в преимущественно черных городских районах и слишком мало в преимущественно белых пригородных районах». В подтверждение своего утверждения The Inclusive Communities Project привел статистические данные, которые показали, что «92,29 % единиц [налогового кредита на жилье с низким доходом] в городе Даллас располагались в переписных районах, где проживало менее 50% жителей Кавказа». Окружной суд, а затем и Апелляционный суд пятого округа США вынесли решение в пользу The Inclusive Communities Project. Департамент жилищного строительства и сообщества Техаса подал апелляцию в Верховный суд США. В деле Техасский департамент жилищного строительства и общественных дел против Inclusive Communities Project, Inc., которое послужило мотивировкой предложенного HUD правила, суд потребовал наличия «причинной связи» между процессом принятия решения и непропорциональными результатами в данных, связанных с расой. Кроме того, один из судей – судья Кеннеди также постановил, что жилищные органы и частные застройщики должны иметь возможность защищаться от разрозненных исков о дискриминации, заявляя и объясняя «действительный интерес, которому служит их политика»³.

Тем самым судьи продемонстрировали, что наличие антидискриминационного и антиконкурентного законодательства в новых условиях не всегда приводит к желаемому эффекту. Кроме того, исследования показали, что такой подход машинного обучения, как «справедливость через неосведомленность», закрепленный последовательно в стандартах и этических принципах в различных странах, также не приводит к желаемым результатам.

¹ Nasdaq. Shirking: Financial terms. URL: <https://www.nasdaq.com/glossary/s/shirking>. Glosary. On-line.

² URL: https://en.wikipedia.org/wiki/Texas_Department_of_Housing_and_Community_Affairs_v._Inclusive_Communities_Project,_Inc.#cite_note-17.

³ URL: https://en.wikipedia.org/wiki/Texas_Department_of_Housing_and_Community_Affairs_v._Inclusive_Communities_Project,_Inc.#cite_note-17.

В контексте ограничения возможностей применения несправедливых алгоритмов искусственного интеллекта представляется важным подчеркнуть и вопросы качества и достоверности анализируемых данных, а также приватности этих данных.

В отчете Европейского парламента о последствиях больших данных для основных прав: конфиденциальности, защиты данных, недискриминации, безопасности и правоприменения¹, Европейский парламент подчеркнул, что «из-за наборов данных и алгоритмических систем, используемых при оценке и прогнозировании на разных этапах обработки данных, большие данные могут приводить не только к нарушениям основных прав человека, но также к дифференцированному обращению и косвенной дискриминации в отношении различных групп людей со схожими характеристиками, особенно в отношении справедливости и равенства возможностей доступа к образованию и занятости, при найме или оценке людей или при определении новых потребительских привычек пользователей социальных сетей»². В отчете содержится призыв к Европейской комиссии, государствам-членам и органам по защите данных «выявить и принять любые возможные меры для минимизации алгоритмической дискриминации и предвзятости и разработать прочную и общую этическую основу для прозрачной обработки персональных данных и автоматизированного принятия решений, которые могут направлять использование данных и постоянное соблюдение законодательства Союза»³.

Качество данных как правовое понятие закреплено в статье 5 Конвенции о защите физических лиц при автоматизированной обработке персональных данных⁴ и относятся к персональным данным, подвергающимся автоматизированной обработке.

Качественные персональные данные:

а) собираются и обрабатываются на справедливой и законной основе;

б) хранятся для определенных и законных целей и не используются иным образом, несовместимым с этими целями;

с) являются адекватными, относящимися к делу и не чрезмерными для целей их хранения;

д) являются точными и, когда это необходимо, обновляются;

е) сохраняются в форме, позволяющей идентифицировать субъекты данных, не дольше, чем это требуется для целей хранения этих данных.

Следует обратить внимание, что качественными могут быть не только персональные данные. В силу статьи 7 Федерального закона «Об информации, информационных технологиях и о защите информации» вводится понятие общедоступной информации – это общеизвестные сведения и иная информация, доступ к которой не ограничен. Общедоступной является любая информация, доступ к которой не ограничен каким-либо законным образом [2, с. 73]. Информация, размещаемая в сети Интернет, которая может быть обработана автоматизированными средствами для повторного ее использования, представляет собой общедоступную информацию, для которой установлен правовой режим «открытых данных».

Повышение доступности и качества данных, необходимых для развития технологий искусственного интеллекта, заявлено как одна из целей Национальной стратегии развития искусственного интеллекта на период до 2030 года⁵: «Основными факторами развития технологий искусственного интеллекта являются увеличение объема доступных данных, в том числе данных, прошедших разметку и структурирование, и развитие информационно-коммуникационной инфраструктуры для обеспечения доступа к наборам таких данных» (п. 37).

Основными направлениями повышения доступности и качества данных, необходимых

¹ European Parliament (2017) Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

² European Parliament (2017) Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

³ European Parliament (2017) Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

⁴ Конвенция о защите физических лиц при автоматизированной обработке персональных данных: заключена в г. Страсбурге 28 янв. 1981 г. [Электронный ресурс]. Доступ из справ.-правовой системы «КонсультантПлюс».

⁵ О развитии искусственного интеллекта в Российской Федерации (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года») [Электронный ресурс]: указ Президента Рос. Федерации от 10 окт. 2019 г. № 490. Доступ из справ.-правовой системы «КонсультантПлюс».

для развития технологий искусственного интеллекта, в соответствии с требованиями законодательства Российской Федерации являются:

а) разработка унифицированных и обновляемых методологий описания, сбора и разметки данных, а также механизма контроля за соблюдением указанных методологий;

б) создание и развитие информационно-коммуникационной инфраструктуры для обеспечения доступа к наборам данных посредством:

– создания (модернизации) общедоступных платформ для хранения наборов данных, соответствующих методологиям описания, сбора и разметки данных;

– хранения наборов данных (в том числе звуковых, речевых, медицинских, метеорологических, промышленных данных и данных систем видеонаблюдения) на общедоступных платформах для обеспечения потребностей организаций – разработчиков в области искусственного интеллекта;

– установления приоритетного доступа российских государственных органов и организаций к общедоступным платформам (п. 38 Стратегии 2030).

Как видим, основной упор сделан на повышение требований к качеству данных, подтвержденных операторами цифровых платформ. В целом российский законодатель идет по пути создания правил и требований для операторов цифровых платформ, собирающих и обрабатывающих данные разного рода. При создании законодательства в сфере сбора и обработки данных в России необходимо учесть, что предотвратить проблему дискриминации разного рода групп населения можно при установлении не только прямого запрета расизма, например, но и путем выработки правил анонимизации общедоступных данных [3].

Возвращаясь к требованиям, адресованным разработчикам искусственного интеллекта и цифровым платформам, стоит привести еще один пример. Специалисты лондонской компании DeepMind предложили в качестве защиты от влияния человеческих предубеждений использовать метод «гипотетическая справедливость» (*counterfactual fairness* [18]). Чтобы сформулировать справедливое и непредвзятое суждение о гражданине, искусственный интеллект формирует гипотетическую ситуацию,

в которой данный гражданин обладает противоположными признаками: женщина превращается в мужчину, бедный – в богатого, афроамериканец – в белого и т.п. Таким образом, реальный статус не влияет на оценку деяний гражданина. Суждение формируется в гипотетической ситуации. Такое суждение, по мнению разработчиков, считается свободным от предубеждений, а значит, справедливым¹.

На наш взгляд, этот кейс есть лучшее свидетельство того, что добиться непредвзятости, смягчить риски применения алгоритмов искусственного интеллекта невозможно только силами специалистов одной какой-то сферы. Требуется объединение усилий программистов, правоведов, философов, социологов для постоянного контроля за новыми проявлениями алгоритмической предвзятости.

Выводы

Если оставить без внимания предвзятые алгоритмы, они могут привести к решениям, которые будут иметь коллективное разрозненное влияние на определенные группы людей даже без намерения программиста проводить различие. Изучать предполагаемые и непредвиденные последствия применения алгоритмов искусственного интеллекта необходимо своевременно, особенно потому, что текущая государственная политика может быть недостаточной для выявления, смягчения и устранения последствий такой неочевидной предвзятости для участников правовых отношений. Решение проблем предвзятости алгоритмов исключительно с помощью технических средств не приведет к желаемым результатам. Мировое сообщество задумалось о введении стандартизации и разработке этических принципов, позволяющих установить рамки для справедливого применения искусственного интеллекта в принятии решений. *Необходимо создание специальных норм, устанавливающих ограничения для алгоритмической предвзятости.* Независимо от сфер выявления таких нарушений они обладают общими признаками недобросовестного поведения участников общественных отношений и могут быть квалифицированы как нарушения

¹ *Этика и «цифра»*: этические проблемы цифровых технологий Аналитический доклад. 2021. URL: <https://ethics.cdto.ranepa.ru/>.

прав человека или нарушения добросовестной конкуренции. Минимизации предвзятости алгоритмов будет способствовать обязательное введение в оборот данных в том виде, который не позволяет проводить явную или скрытую сегрегацию различных групп населения. В то же время подлежат анализу не просто данные без явных признаков групп, но обязательно в полном их разнообразии. Общеприменимая модель искусственного интеллекта тем самым будет строиться на анализе данных всех социально-правовых групп общества.

Библиографический список

1. Вейценбаум Дж. Возможности вычислительных машин и человеческий разум: От суждений к вычислениям. М.: Радио и связь, 1982. 369 с.
2. Савельев А. И. Комментарий к Федеральному закону от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и защите информации» (постатейный). М.: Статут, 2015. 320 с.
3. Харитонова А. Р. Сохранность и анонимность персональных данных в социальных сетях // Предпринимательское право. 2019. № 4. С. 48 – 55. [Приложение «Право и Бизнес»].
4. Харитонова Ю. С. Контекстная (поведенческая) реклама и право: точки пересечения // Рожкова М. А. Право в сфере Интернета: сб. ст. / М. З. Али, Д. В. Афанасьев, В. А. Белов [и др.]; рук. авт. кол. и отв. ред. М. А. Рожкова. М.: Статут, 2018. 528 с.
5. Харитонова Ю. С., Савина В. С. Технология искусственного интеллекта и право: вызовы современности // Вестник Пермского университета. Юридические науки. 2020. Вып. 3. С. 524–549. DOI: 10.17072/1995-4190-2020-49-524-549.
6. Adelman L. Unnatural Causes: Is Inequality Making Us Sick? // Preventing Chronic Disease. 2007. Vol. 4., № 4.
7. Agniel D., Kohane I. K., Weber G. M. Biases in Electronic Health Record Data due to Processes Within the Healthcare System: Retrospective Observational Study // BMJ, 2018. Vol. 361, № 8151. k1479.
8. Angwin J., Larson J, Mattu S, Kirchner L., Machine Bias 2016. May 23, 2016. URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
9. Baeza-Yates, R. Data and Algorithmic Bias in the Web. // Proceedings of the 8th ACM Conference on Web Science. 2016, May.
10. Bakshy E., Messing S., Adamic L. A. Exposure to Ideologically Diverse News and Opinion on Facebook // Science. 2015. № 348(6239). Pp.1130–1132. DOI: 10.1126/science.aaa1160.
11. Barr A. Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms // The Wall Street Journal. 2015. № 1. URL: <https://www.wsj.com/articles/BL-DGB-42522>.
12. Bartolini C., Lenzini G., Santos C. An Agile Approach to Validate a Formal Representation of the GDPR // JSAI: Annual Conference of the Japanese Society for Artificial Intelligence. Springer, Cham. 2018. November. Pp. 160–176.
13. Bengio Y. Learning Deep Architectures for AI. Now Publishers Inc. 2009.
14. Benkler Y. Don’t Let Industry Write the Rules for AI // Nature, 2019. № 569(7754). Pp. 161–162. DOI: <https://doi.org/10.1038/d41586-019-01413-1>.
15. Buolamwini J., Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Proceedings of the Conference on Fairness, Accountability and Transparency. PMLR. 2018. January. Vol. 81. Pp. 77–91. URL: <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
16. Burns N. Why Should We Expect Algorithms to Be Biased // MIT Technology Review 2016. URL: <https://www.technologyreview.com/s/601775/why-we-should-expect-algorithms-to-be-biased/>.
17. Casacuberta D. Bias in a Feedback Loop: Fuelling Algorithmic Injustice // CCCB LAB. May 9, 2018.
18. Chiappa S. Path-Specific Counterfactual Fairness// Proceedings of the AAAI Conference on Artificial Intelligence. 2019. July. Vol. 33, № 1. Pp. 7801–7808.
19. Colleoni E., Rozza A., Arvidsson A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data // Journal of Communication. 2014. № 64(2). Pp. 317–332.
20. Courtland R. Bias Detectives: the Researchers Striving to Make Algorithms Fair // Nature. 2018. № 558. Pp. 357–60.
21. Del Vicario M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G. et al. The spreading of misinformation online // Proceedings

of the National Academy of Sciences. 2016. № 113(3). Pp. 554–559.

22. *Dixon B.* What is Algorithmic Bias? // *TechTalks*. March 26, 2018. URL: <https://bdtech-talks.com/2018/03/26/racist-secsist-ai-depp-leaning-algoritms>.

23. *Edizel B., Bonchi F., Hajian S., Panisson A., Tassa T.* FaiRecSys: Mitigating Algorithmic Bias in Recommender Systems. *International Journal of Data // Science and Analytics*. 2020. № 9(2). Pp. 197–213.

24. *Fitzpatrick T. B.* «Soleil et peau» [Sun and skin] // *Journal de Médecine Esthétique* (in French). 1975. № (2). Pp. 33–34.

25. *Floridi L.* Translating principles into Practices of Digital Ethics: Five Risks of Being Unethical // *Philosophy & Technology*. 2019. № 32(2). Pp. 185–193.

26. *Garcia M.* Racist in the Machine: The disturbing Implications of Algorithmic Bias // *World Policy Journal*. 2016. № 33(4). Pp. 111–117.

27. *Garimella K., De Francisci Morales G., Gionis A., Mathioudakis M.* Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. // *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2018. P. 913–922.

28. *Hacker P.* Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination under EU Law // *Common Market Law Review*. 2018. № 55(4).

29. *Hagendorff T.* The Ethics of AI Ethics: an Evaluation of Guidelines // *Minds and Machines*. 2020. № 30(1). Pp. 99–120.

30. *Hajian S., Bonchi F., Castillo C.* Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Minin // *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. Pp. 2125–2126.

31. *Hao K.* This is How AI Bias Really Happens – And Why It’s so Hard to Fix // *MIT Technology Review*. URL: <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/>.

32. *Heidari H., Nanda V., Gummadi K. P.* On the Long-Term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. URL: <https://arxiv.org/abs/1903.01209>.

33. *Ignatieff M.* Political Polarization in the American Public. *Annual Colloquium on Fundamental Rights*. Brussels, 2016.

34. *Jackson J. R.* Algorithmic Bias // *Journal of Leadership, Accountability & Ethics*. 2018. № 15(4). Pp. 55–65.

35. *Jobin A., Ienca M., Vayena E.* The Global Landscape of AI Ethics Guidelines // *Nature Machine Intelligence*. 2019. № 1(9). Pp. 389–399.

36. *Klingenberg C. O., Borges M. A. V., Antunes Jr, J. A. V.* Industry 4.0 as a Data-Driven Paradigm: A Systematic Literature Review on Technologies // *Journal of Manufacturing Technology Management*. 2019. Vol. 32. Issue 3. Pp. 570–592.

37. *Koene A., Douthwaite L., Seth S.* IEEE P7003™ Standard for Algorithmic Bias Considerations: Work in Progress Paper // *Proceedings of the International Workshop on Software Fairness*. 2018. May. Pp. 38–41.

38. *Koren Y.* Collaborative Filtering with Temporal Dynamics // *Communications of the ACM*. 2010. № 53(4). Pp. 89–97.

39. *Lipton Z. C.* The Foundations of Algorithmic Bias. 2016. URL: <http://approximately-correct.com/2016/11/07/the-foundations-of-algorithmic-bias/>.

40. *Ludbrook F., Michalikova K. F., Musova Z., Suler P.* Business Models for Sustainable Innovation in Industry 4.0: Smart Manufacturing Processes, Digitalization of Production Systems, and Data-Driven Decision Making // *Journal of Self-Governance and Management Economics*. 2019. № 7(3). Pp. 21–26.

41. *Mäs M., Bischofberger L.* Will the Personalization of Online Social Networks Foster Opinion Polarization? 2015. URL: <http://dx.doi.org/10.2139/ssrn.2553436>.

42. *Moutafis R.* We’re Facing a Fake Science Crisis, and AI is Making it Worse // *Journals are Retracting More and More Papers Because They’re not by the Authors They Claim to Be*. June 8, 2021 Updated: June 9, 2021. URL: <https://builtin.com/artificial-intelligence/ai-fake-science>.

43. *Panch T., Szolovits P., Atun R. J.* Artificial Intelligence, Machine Learning and Health Systems // *Journal of Global Health*. 2018 Dec.; № 8(2):020303. DOi: 10.7189/jogh.08.020303

44. *Panch T., Mattie H., Atun R.* Artificial Intelligence and Algorithmic Bias // *Implications*

for Health Systems. // Journal of Global Health. 2019. № 9(2). DOI: 10.7189/jogh.09.020318.

45. Pariser E. The filter bubble: What the Internet is Hiding from You. Penguin UK, 2011.

46. Paulus J. K., Kent D. M. Predictably Unequal: Understanding and Addressing Concerns that Algorithmic Clinical Prediction May Increase Health Disparities // NPJ Digital Medicine. 2020. № 3(1). Pp. 1–8.

47. Rodrigues R. Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities // Journal of Responsible Technology. 2020. Vol. 4, 100005. URL: <https://doi.org/10.1016/j.jrt.2020.100005>.

48. Schmidt A. L., Zollo F., Del Vicario M., Bessi A., Scala A., Caldarelli G., et al. Anatomy of News Consumption on Facebook // Proceedings of the National Academy of Sciences. 2017. № 114(12). Pp. 3035–3039.

49. Schmidt A.L., Zollo F., Scala A., Quattrocioni W. Polarization Rank: A Study on European News Consumption on Facebook. arXiv preprint arXiv:180508030. 2018.

50. Schroeder D., Chatfield K., Singh M., Chennells R., Herisson-Kelly P. Ethics Dumping and the Need for a Global Code of Conduct // Equitable Research Partnerships, Springer. 2019.

51. Shaulova T. Artificial Intelligence vs. Gender Equality // International Relations and Dialogue of Cultures. 2019. № 7. Pp. 52–54.

52. Simonite T. When it Comes to Gorillas, Google Photos Remains Blind. Wired, 2018. January 11. URL: <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.

53. Sirbu A., Pedreschi D., Giannotti F., Kertész J. Algorithmic Bias Amplifies Opinion Fragmentation and Polarization: A Bounded Confidence Model // PloS one. 2019. № 14(3). <https://doi.org/10.1371/journal.pone.0213246>. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213246>.

54. Smith L. Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making // Future of Privacy Forum. 2017. URL: <https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making>.

55. Stojanovic L., Dinic M., Stojanovic N., Stojadinovic A. Big-Data-Driven Anomaly Detection in Industry (4.0): An Approach and a Case Study // 2016 IEEE International Conference on

Big Data (Big Data) (2016, December. Pp. 1647–1652). IEEE.

56. Strickland E. Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients // IEEE Spectrum. October 24, 2019.

57. Wael W. D., Artificial Intelligence SC 42 Overview, ITU Workshop on AI and Data Commons January. 2020. Geneva, Switzerland. URL: https://www.itu.int/en/ITU-T/extcoop/ai-data-commons/Documents/ISO_IEC%20JTC1%20SC%2042%20Keynote_Wael%20Diab.pdf.

58. Winfield A. An Updated Round Up of Ethical Principles of Robotics and AI. Retrieved on August 13th, 2019. URL: <https://tex.stackexchange.com/questions/3587/how-can-i-use-bibtex-to-cite-a-web-page>.

References

1. Weizenbaum J. *Vozmozhnosti vychislitel'nykh mashin i chelovecheskiy razum: Ot suzheniy k vychisleniyam* [Computer Power and Human Reason. From Judgment to Calculation]. Moscow, 1982. 369 p. (In Russ.).

2. Savel'ev A. I. *Kommentariy k Federal'nomu zakonu ot 27 iyulya 2006 g. № 149-FZ "Ob informatsii, informatsionnykh tekhnologiyakh i zashhite informatsii" (postateyny)* [Commentary on the Federal Law No. 149-FZ of July 27, 2006 'On Information, Information Technologies and Information Protection' (Article-by-Article)]. Moscow, 2015. 320 p. (In Russ.).

3. Kharitonova A. R. *Sokhrannost' i anonimnost' personal'nykh dannykh v sotsial'nykh setyakh* [Security and Anonymity of Personal Data in Social Media]. *Predprinimatel'skoe pravo. Prilozhenie "Pravo i Biznes"*- Entrepreneurial Law. 'Law and Business' Supplement. 2019. Issue 4. Pp. 48–55. (In Russ.).

4. Kharitonova Yu. S. *Kontekstnaya (povedencheskaya) reklama i pravo: tochki peresecheniya* [Contextual (Behavioral) Advertising and Law: Points of Intersection]. *Pravo v sfere Internet: sbornik statey* [Law in the Internet Sphere: Collection of Articles]; M. Z. Ali, D. V. Afanas'ev, V. A. Belov et al.; ed. by M. A. Rozhkova. Moscow, 2018. 528 p. (In Russ.).

5. Kharitonova Yu. S., Savina V. S. *Tekhnologiya iskusstvennogo intellekta i pravo: vyzovy sovremennosti* [Artificial Intelligence Technology and Law: Challenges of Our Time]. *Vestnik Permskogo universiteta. Juridicheskie nauki* –

Perm University Herald. Juridical Sciences. 2020. Issue 3. Pp. 524–549. DOI: 10.17072/1995-4190-2020-49-524-549. (In Russ.).

6. *Adelman L.* Unnatural Causes: Is Inequality Making Us Sick? *Preventing Chronic Disease*. 2007. Vol. 4. Issue 4. (In Eng.).

7. *Agniel D., Kohane I. K., Weber G. M.* Biases in Electronic Health Record Data due to Processes within the Healthcare System: Retrospective Observational Study. *BMJ*. 2018. Vol. 361. Issue 8151. k1479. (In Eng.).

8. *Angwin J., Larson J., Mattu S., Kirchner L.* Machine Bias. May 23, 2016. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. (In Eng.).

9. *Baeza-Yates R.* Data and Algorithmic Bias in the Web. *Proceedings of the 8th ACM Conference on Web Science*. May 2016. (In Eng.).

10. *Bakshy E., Messing S., Adamic L. A.* Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*. 2015. Vol. 348. Issue 6239. Pp. 1130–1132. DOI: 10.1126/science.aaa1160. (In Eng.).

11. *Barr A.* Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms. *The Wall Street Journal*. July 1, 2015. Available at: <https://www.wsj.com/articles/BL-DGB-42522>. (In Eng.).

12. *Bartolini C., Lenzini G., Santos C.* An Agile Approach to Validate a Formal Representation of the GDPR. *JSAI: Annual Conference of the Japanese Society for Artificial Intelligence* Ed. by K. Kojima, M. Sakamoto, K. Mineshima, K. Satoh. Springer, Cham, 2019. Pp. 160–176. (In Eng.).

13. *Bengio Y.* Learning Deep Architectures for AI. Now Publishers Inc., 2009. (In Eng.).

14. *Benkler Y.* Don’t Let Industry Write the Rules for AI. *Nature*. 2019. Issue 569(7754). Pp. 161–162. DOI: <https://doi.org/10.1038/d41586-019-01413-1>. (In Eng.).

15. *Buolamwini J., Gebru T.* Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR. 2018. Vol. 81. Pp. 77–91. (In Eng.).

16. *Burns N.* Why Should We Expect Algorithms to Be Biased. *MIT Technology Review*. June 24, 2016. Available at: <https://www.technologyreview.com/s/601775/why-we-should-expect-algorithms-to-be-biased/>. (In Eng.).

17. *Casacuberta D.* Bias in a Feedback Loop: Fuelling Algorithmic Injustice. *CCCB LAB*. May 9, 2018. (In Eng.).

18. *Chiappa S.* Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019. Vol. 33. Issue 1. Pp. 7801–7808. (In Eng.).

19. *Colleoni E., Rozza A., Arvidsson A.* Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*. 2014. Issue 64(2). Pp. 317–332. (In Eng.).

20. *Courtland R.* Bias Detectives: the Researchers Striving to Make Algorithms Fair. *Nature*. 2018. Issue 558(7710). Pp. 357–360. (In Eng.).

21. *Del Vicario M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G. et al.* The Spreading of Misinformation Online. *Proceedings of the National Academy of Sciences*. 2016. Issue 113(3). Pp. 554–559. (In Eng.).

22. *Dixon B.* What is Algorithmic Bias? *TechTalks*. March 26, 2018. Available at: <https://bdtechtalks.com/2018/03/26/racist-secsist-ai-depp-leaning-algorithms>. (In Eng.).

23. *Edizel B., Bonchi F., Hajian S., Panisson A., Tassa T.* FaiRecSys: Mitigating Algorithmic Bias in Recommender Systems. *International Journal of Data Science and Analytics*. 2020. Issue 9(2). Pp. 197–213. (In Eng.).

24. *Fitzpatrick T. B.* ‘Soleil et peau’ [Sun and skin]. *Journal de Médecine Esthétique*. 1975. Issue 2. Pp. 33–34. (In Eng.).

25. *Floridi L.* Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*. 2019. Issue 32(2). Pp. 185–193. (In Eng.).

26. *Garcia M.* Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*. 2016. Issue 33(4). Pp. 111–117. (In Eng.).

27. *Garimella K., De Francisci Morales G., Gionis A., Mathioudakis M.* Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2018. Pp. 913–922. (In Eng.).

28. *Hacker P.* Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law. *Com-*

mon Market Law Review. 2018. Issue 55(4). (In Eng.).

29. Hagendorff T. The Ethics of AI Ethics: an Evaluation of Guidelines. *Minds and Machines*. 2020. Issue 30(1). Pp. 99–120. (In Eng.).

30. Hajian S., Bonchi F., Castillo C. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. Pp. 2125–2126. (In Eng.).

31. Hao K. This Is How AI Bias Really Happens – And Why It's So Hard to Fix. *MIT Technology Review*. February 4, 2019. (In Eng.).

32. Heidari H., Nanda V., Gummadi K. P. On the Long-Term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. Available at: <https://arxiv.org/abs/1903.01209> (In Eng.).

33. Ignatieff M. Political Polarization in the American Public. Annual Colloquium on Fundamental Rights. Brussels, 2016. (In Eng.).

34. Jackson J. R. Algorithmic Bias. *Journal of Leadership, Accountability & Ethics*. 2018. Vol. 15. Issue 4. Pp. 55–65. (In Eng.).

35. Jobin A., Ienca M., Vayena E. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*. 2019. Issue 1(9). Pp. 389–399. (In Eng.).

36. Klingenberg C. O., Borges M. A. V., Antunes Jr J. A. V. Industry 4.0 as a Data-Driven Paradigm: A Systematic Literature Review on Technologies. *Journal of Manufacturing Technology Management*. 2019. Vol. 32. Issue 3. Pp. 570–592. (In Eng.).

37. Koene A., Dowthwaite L., Seth S. IEEE P7003™ Standard for Algorithmic Bias Considerations: Work in Progress Paper. *Proceedings of the International Workshop on Software Fairness*. May 2018. Pp. 38–41. (In Eng.).

38. Koren Y. Collaborative Filtering with Temporal Dynamics. *Communications of the ACM*. 2010. Issue 53(4). Pp. 89–97. (In Eng.).

39. Lipton Z. C. The Foundations of Algorithmic Bias. 2016. (In Eng.). Available at: <http://approximatelycorrect.com/2016/11/07/the-foundations-of-algorithmic-bias/>.

40. Ludbrook F., Michalikova K. F., Musova Z., Suler P. Business Models for Sustainable Innovation in Industry 4.0: Smart Manufacturing Processes, Digitalization of Production Systems,

and Data-Driven Decision Making. *Journal of Self-Governance and Management Economics*. 2019. Issue 7(3). Pp. 21–26. (In Eng.).

41. Mäs M, Bischofberger L. Will the Personalization of Online Social Networks Foster Opinion Polarization? 2015. Available at: <http://dx.doi.org/10.2139/ssrn.2553436>. (In Eng.).

42. Moutafis R. We're Facing a Fake Science Crisis, and AI is Making It Worse: Journals are Retracting More and More Papers Because They're not by the Authors They Claim to Be. June 8, 2021. Updated June 9, 2021. <https://builtin.com/artificial-intelligence/ai-fake-science>. (In Eng.).

43. Panch T., Szolovits P., Atun R. J. Artificial Intelligence, Machine Learning and Health Systems. *Journal of Global Health*. December 2018. Issue 8(2):020303. DOI: 10.7189/jogh.08.020303. (In Eng.).

44. Panch T., Mattie H., Atun R. Artificial Intelligence and Algorithmic Bias: Implications for Health Systems. *Journal of Global Health*. 2019. Issue 9(2). DOI: 10.7189/jogh.09.020318. (In Eng.).

45. Pariser E. The Filter Bubble: What the Internet Is Hiding from You. Penguin UK, 2011. (In Eng.).

46. Paulus J. K., Kent D. M. Predictably Unequal: Understanding and Addressing Concerns that Algorithmic Clinical Prediction May Increase Health Disparities. *NPJ Digital Medicine*. 2020. Issue 3(1). Pp. 1–8. (In Eng.).

47. Rodrigues R. Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities. *Journal of Responsible Technology*. 2020. Vol. 4. 100005. (In Eng.).

48. Schmidt A. L., Zollo F., Del Vicario M., Bessi A., Scala A., Caldarelli G. et al. Anatomy of News Consumption on Facebook. *Proceedings of the National Academy of Sciences*. 2017. Issue 114(12). Pp. 3035–3039. (In Eng.).

49. Schmidt A. L., Zollo F., Scala A., Quattrociocchi W. Polarization Rank: A Study on European News Consumption on Facebook. arXiv preprint arXiv:180508030. 2018. (In Eng.).

50. Schroeder D., Chatfield K., Singh M., Chennells R., Herissone-Kelly P. Ethics Dumping and the Need for a Global Code of Conduct. *Equitable Research Partnerships*. Springer, 2019. (In Eng.).

51. Shaulova T. Artificial Intelligence vs. Gender Equality. *International Relations and Di-*

dialogue of Cultures. 2019. Issue 7. Pp. 52–54. DOI: 10.1870/HUM/2304-9480.7.04. (In Eng.).

52. *Simonite T.* When It Comes to Gorillas, Google Photos Remains Blind. *Wired*. January 11, 2018. DOI 10.1870/HUM/2304-9480.7.04. (In Eng.).

53. *Sîrbu A., Pedreschi D., Giannotti F., Kertész J.* Algorithmic Bias Amplifies Opinion Fragmentation and Polarization: A Bounded Confidence Model. *PloS one*. 2019. Issue 14(3). Available at: <https://doi.org/10.1371/journal.pone.0213246>. (In Eng.).

54. *Smith L.* Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making. *Future of Privacy Forum*. December 11, 2017. Available at: <https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-theharms-of-automated-decision-making>. (In Eng.).

55. *Stojanovic L., Dinic M., Stojanovic N., Stojadinovic A.* Big-Data-Driven Anomaly Detec-

tion in Industry (4.0): An Approach and a Case Study. 2016 IEEE International Conference on Big Data (Big Data). 2016. Pp. 1647–1652.

56. *Strickland E.* Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients. *IEEE Spectrum*. October 24, 2019. (In Eng.).

57. *Wael W. D.* Artificial Intelligence. SC 42 Overview, ITU Workshop on AI and Data Commons. Geneva, Switzerland, January 2020. Available at: https://www.itu.int/en/ITU-T/ext-coop/ai-data-commons/Documents/ISO_IEC%20JTC1%20SC%2042%20Keynote_Wael%20Diab.pdf. (In Eng.).

58. *Winfield A.* An Updated Round Up of Ethical Principles of Robotics and AI. April 18, 2019. Available at: <https://tex.stackexchange.com/questions/3587/how-can-i-use-bibtex-to-cite-a-web-page>. (In Eng.).

Информация об авторах:

Ю. С. Харитонов, доктор юридических наук, профессор, профессор кафедры предпринимательского права

Руководитель НОЦ «Центр правовых исследований искусственного интеллекта и цифровой экономики»

Московский государственный университет имени М. В. Ломоносова

119991, Россия г. Москва, Ленинские горы, 1

ORCID: 0000-0001-7622-6215

ResearcherID: K-7495-2016

Статьи в БД «Scopus» / «Web of Science»:

DOI: 10.17072/1995-4190-2019-43-121-145

DOI: 10.17072/1995-4190-2020-49-524-549

DOI: 10.17072/1995-4190-2016-34-451-460

В. С. Савина, доктор юридических наук, доцент, профессор кафедры гражданско-правовых дисциплин

Российский экономический университет

имени Г.В. Плеханова

117997, Россия, г. Москва, Стремянный пер., 36

ORCID: 0000-0002-8385-9421

ResearcherID: G-2782-2014

Статьи в БД «Scopus» / «Web of Science»:

DOI: 10.17072/1995-4190-2020-49-524-549

Ф. Паньини, технический инженер

LOYTEC Electronics GmbH

1170, Австрия, г. Вена, Блюменгассе, 35

ORCID: 0000-0003-4618-0740

ResearcherID: AAU-6991-2021

About the authors:

Yu. S. Kharitonova

Lomonosov Moscow State University

1, Leninskie Gory, Moscow, 119991

ORCID: 0000-0001-7622-6215

ResearcherID: K-7495-2016

Articles in Scopus / Web of Science:

DOI: 10.17072/1995-4190-2019-43-121-145

DOI: 10.17072/1995-4190-2020-49-524-549

DOI: 10.17072/1995-4190-2016-34-451-460

V. S. Savina

Plekhanov Russian University of Economics

36, Stremyanniy lane, Moscow, 117997, Russia

ORCID: 0000-0002-8385-9421

ResearcherID: G-2782-2014

Articles in Scopus / Web of Science:

DOI: 10.17072/1995-4190-2020-49-524-549

F. Pagnini

LOYTEC Electronics GmbH

35, Blumengasse, Vienna, 1170, Austria

ORCID: 0000-0003-4618-0740

ResearcherID: AAU-6991-2021