УДК 338.49
doi:10.18720/SPBPU/2/id23-77

Качалов Роман Михайлович [1],
глав. науч. сотр., д-р экон. наук, профессор;
Слепцова Юлия Анатольевна [2],
ведущ. науч. сотр., канд. экон. наук, доцент

ЭТИЧЕСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В УПРАВЛЕНИИ ПРЕДПРИЯТИЕМ

[1, 2] Россия, Москва,
Центральный экономико-математический институт РАН,
[1] kachalov1ya@yandex.ru, [2] julia_sleptsova@mail.ru

Аннотация. Алгоритмы искусственного интеллекта уже внесли значительный прогресс в таких сферах, как медицина, транспорт и финансы, в настоящее время необходимо понять какие побочные эффекты искусственного интеллекта могут вызвать после их интеграции в производство продуктов и услуг. Целью настоящей работы является исследование этических аспектов выявления факторов риска неустойчивости алгоритмов искусственного интеллекта. В работе идентифицированы факторы риска обучения алгоритмов искусственного интеллекта, предложены принципы процедур сертификации таких алгоритмов.

Ключевые слова: этические аспекты, алгоритмы, искусственный интеллект, факторы риска, антирисковые управленческие воздействия, операциональная теория управления уровнем риска.

Roman M. Kachalov [1],
Head of Laboratory of Publishing and Marketing Activity,
Doctor of Sciences (Economics), Professor;
Yulia A. Sleptsova [2],
Leading Researcher, Laboratory of Publishing and Marketing Activity,
PhD (Economics), Associate Professor

ETHICAL PROBLEMS OF USING ARTIFICIAL INTELLIGENCE ALGORITHMS IN ENTERPRISE MANAGEMENT

[1, 2] Central Economics and Mathematics Institute of
Russian Academy of Sciences, Moscow, Russia,
[1] kachalov1ya@yandex.ru, [2] julia_sleptsova@mail.ru

Abstract. Artificial intelligence algorithms have already made significant progress in such areas as medicine, transport and finance, now it is necessary to understand what negative effects artificial intelligence can cause after their integration into the production of products and services. The purpose of this work is to study the ethical aspects of identifying risk factors for the instability of artificial intelligence algorithms. The paper identifies risk factors for learning artificial intelligence algorithms, and suggests the principles of certification procedures for such algorithms.

Keywords: ethical aspects, algorithms, artificial intelligence, risk factors, anti-risk managerial impacts, operational theory of risk management.

Введение

В октябре 2019 года была утверждена «Национальная стратегия развития искусственного интеллекта на период до 2030 года»[1], основные задачи которой были сформулированы как «… обеспечение роста благосостояния и качества жизни ее населения, обеспечение национальной безопасности и правопорядка, достижение устойчивой конкурентоспособности российской экономики, в том числе лидирующих позиций в мире в области искусственного интеллекта». Стратегия предусматривает ряд приоритетных направлений развития теории искусственного интеллекта, которые можно условно разделить на две группы: экономические и социальные. К экономическим относятся задачи повышения точности прогнозирования; автоматизации рутинных процессов; использования автономных устройств; повышения лояльности потребителей; оптимизации процессов подбора и обучения кадров. Среди социальных целей выделены такие задачи как повышение качества здравоохранения, образования, государственных и муниципальных услуг, снижение затрат на их предоставление.

[1] Указ Президента Российской Федерации № 490 от 10 октября 2019 года «О развитии искусственного интеллекта в Российской Федерации».
http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf (дата обращения: 03.10.2022).

Стратегия также выделяет некоторое подмножество задач, которые необходимо решить для успешного совершенствования алгоритмов искусственного интеллекта: создание высокопроизводительных рабочих мест; обеспечение конкурентоспособных условий труда для специалистов в сфере искусственного интеллекта; создание стимулов для развития науки и исследований в области нейронных сетей; формирование комплексной системы безопасности при создании, развитии, внедрении и использовании алгоритмов искусственного интеллекта.

По многим вопросам алгоритмы *искусственного интеллекта* (ИИ), находящиеся в стадии разработки, далеки от выполнения минимальных требований безопасности, которых можно ожидать от применения подобных систем [4]. Алгоритмы ИИ могут также быть подвержены различным факторам риска, связанным с их надежностью. Так, например, отсутствие тестирования в критических ситуациях порождает множество факторов риска в области защиты и конфиденциальности данных, используемых для моделей машинного обучения и алгоритмов ИИ [7]. На данный момент можно признать перспективными несколько путей разработки и реализации стандартов и сертификации безопасности и надежности алгоритмов ИИ, встроенных в реальные системы. Эти траектории, как свидетельствует анализ практики включают в себя такие положения, как, например, нижеследующие:

– разработка методики выявления и оценки факторов риска, связанных с использованием алгоритмов ИИ индивидуальными пользователями и предприятиями;

– введение в практику стандартизированных тестов для оценки надежности алгоритмов ИИ, в отношении данных, которые использованы для обучения;

– обеспечение прозрачности концепции моделирования машинного обучения и дизайна интерпретаций алгоритмов ИИ для устранения потенциально негативных последствий.

Целью настоящей работы является исследование этических аспектов и подходов к построению типологии факторов риска неустойчивости и слабой интерпретируемости алгоритмов ИИ, включая технологии прогнозирования поведения сотрудников современных предприятий в критических ситуациях в условиях применения алгоритмов ИИ.

1. Методология исследования

Алгоритмы ИИ включают в себя множество математических методов оптимизации и машинного обучения, которые позволяют имитировать когнитивные функции человека. Операциональная теория управления риском должна быть дополнена элементами культуры управления риском, в которые должны быть заложены, в том числе принципы трудовой этики.

В данный момент существенными становятся проблемы разработки этических норм не только на этапе использования искусственных интеллектуальных систем, но и на этапе разработки программного обеспечения, робототехники и т. д. В качестве основных этических проблем, требующих своего разрешения в данной области можно выделить такие задачи, как: определение этических принципов и стандартов, регламентирующих разработку и применение алгоритмов ИИ, а также фундаментальные представления о том, как должны разрабатываться и использоваться данные технологии с учетом их потенциального влияния на все сферы жизни общества.

Известно, что многие общественные, политические, экономические и технологические организации, такие как ЮНЕСКО, Институт инженеров электротехники и электроники (IEEE), а также правительства крупнейших мировых государств, в настоящее время активно занимаются разработкой этических норм и правил в области применения алгоритмов ИИ. Однако, данные документы и рекомендации касаются в основном разработки и применения искусственных интеллектуальных систем, тогда как проблемам взаимоотношений работников и программных продуктов, роботов и IT-продуктов в рамках предприятий и организаций не уделяется должного внимания [3].

При разработке алгоритмов ИИ в основном преобладают техники машинного обучения, главной особенностью которых является построение системы рассуждений непосредственно из большого массива данных, без явных правил для генерации результата процесса [2].

Универсальность этих алгоритмов делает их очень привлекательными для широкого спектра применений. Кроме того, исследователи машинного обучения с самого начала приняли открытый подход к сотрудничеству и распространению большого набора ресурсов, от программного обеспечения до наборов данных, документации, свободно доступных каждому. Этот подход способствовал росту популярности машинного обучения в научных и инженерных сообществах и использованию преимуществ огромных объемов данных, собранных в цифровых системах. С другой стороны, со стороны пользователей возникли конкретные требования предоставлять понятные разъяснения при автоматизированном принятии решений без участия человека.

Машинное обучение состоит из набора математических методов, сформулированных на базе сочетания нескольких алгоритмических языков, теории статистического обучения и оптимизации, целью которой является извлечение информации из множества данных, которые включают в себя изображения, записи датчиков и текстов для решения проблем, связанных с этими данными, такими, как классификация, распознавание, генерация и т. д. Это многоступенчатый процесс, который

должен начинаться с определения целей конкретного алгоритма или целой алгоритмической системы, ее функционала, а продолжаться тестированием в условиях, максимально приближенных к реальной экономической деятельности. Проблема формализации этических норм включает в себя два основных типа задач. Во-первых, это создание формализованных представлений таких норм, а во-вторых, выбор соответствующего математического аппарата для работы с этими нормами, то есть задачи сопоставления, измерения, анализа и т. д. [1].

2. Результаты исследования

Под алгоритмами ИИ в настоящем исследовании будем понимать набор алгоритмов, способных к рациональному, автономному принятию решений и адаптации к сложной среде, и к ранее неизвестным обстоятельствам. Применение инструментов ИИ предполагает, что алгоритмические системы смогут самостоятельно принимать во внимание этические соображения. Однако, фактически только разработчики и люди, формирующие массивы данных для машинного обучения алгоритмов ИИ, определяют, как работают алгоритмические системы, поэтому необходимо обсуждать этические предпосылки и последствия такого выбора.

2.1. Факторы риска обучения алгоритмов ИИ

Алгоритмы ИИ по своей природе более сложны, чем классические системы принятия решений из-за нелинейной системы обратных связей между алгоритмами и наборами данных, которые в совокупности составляют модели обучения, действующие как реальные системы рассуждений, разрабатывающие прогнозы на основе входных данных. Эти модели затем встраиваются в более традиционные программы, часто в сочетании с другими частями программного обеспечения, архитектуры, которых реализуются с использованием инструментов программирования. Если эти модели вводятся некорректно или небрежно, то целостность всей архитектуры системы может оказаться под угрозой разрушения, так как контроль безопасности перестает быть адекватным решаемым задачам.

В качестве антирисковых управленческих воздействий в таких ситуациях необходимы внешние проверки, независимо от фазы обучения алгоритма ИИ. При этом желательно не допускать переопределение алгоритмов, поскольку такие алгоритмы не запоминают какой-либо значимый паттерн, а запоминают только входные данные, значительно снижая мощность обобщения. Такие внешние проверки могут быть в некоторой степени сравнимы с этапом тестирования в процедурах машинного обучения, в котором качество модели дополнительно проверяется на ранее неиспользовавшихся и неизвестных данных. Такие антирисковые управленческие воздействия направлены на минимизацию уровня риска смещения спектра входных данных конкретных алгоритмов ИИ.

В этом случае адекватные факторы риска могут быть выявлены при наличии примеров в наборе данных, которые отражают разнообразие и сложность реальных ситуаций, то есть надо иметь ввиду, что наборы данных для обучения могут не отражать реальных ситуаций риска, выявляемые эмпирическим или экспертным способом.

Факторы умышленного повреждения набора данных могут быть идентифицированы при намеренном введении ложных данных на этапе обучения алгоритмов ИИ. Такие повреждения могут быть произведены с целью снижения производительности алгоритмов ИИ или для умышленного, незаметного введения вредоносных элементов, которые могут быть использованы впоследствии.

Умышленное повреждение данных становится возможным при наличии способности алгоритмов ИИ приобретать новые паттерны путем постоянной переподготовки в режиме реального времени с использованием вновь получаемых данных. Такого рода дизайн открывает возможность для злоумышленников постепенно вводить такие данные, которые будут постепенно смещать границы решений алгоритмов ИИ.

2.2. Стандарты и процедуры сертификация алгоритмов ИИ

Для создания благоприятной экосистемы вокруг новых технологий, которые будут гарантировать соответствие направлений использования алгоритмов ИИ гуманистическим ценностям, могут быть использованы такие фундаментальные компоненты, как стандарты и процедуры сертификации. Они, с одной стороны, смогут позволить предприятиям в промышленных отраслях делиться передовым опытом, поощряющим совместимость и интеграцию алгоритмов ИИ в существующую инфраструктуру. А, с другой стороны, надзорным органам вырабатывать эффективную политику, защищающую права граждан, а пользователям понимать и доверять новинкам и, в то же время, возможным сбоям алгоритмов ИИ.

Соответствующее регулирование должно обеспечивать безопасность пользователей и систем в течение всего жизненного цикла алгоритмов ИИ. Это означает, что безопасными должны быть методы разработки программного обеспечения, сертификации, аудита и контроля, реализация которых должна расширять текущую практику компьютерной безопасности. Схема сертификации алгоритмов ИИ может быть основана на выявлении и скрупулезной оценке факторов риска воздействия алгоритмов ИИ на окружающую среду, а также факторов риска уязвимости существующих алгоритмов со стороны недобросовестных сотрудников или третьих лиц. Поэтому должно проводиться обширное тестирование алгоритмов ИИ, оценка их прозрачности и эффективности.

3. Ограничения исследования

Текущие технические ограничения алгоритмов искусственного интеллекта могут вызвать трудности в практической реализации соответствующих данных для субъекта прав. Одно из этих ограничений относится к степени интерпретируемости алгоритмов искусственного интеллекта, которая связана с правом субъектов-генераторов данных получать значимую информацию о логике, заложенной в автоматические процессы принятия решений, и право субъектов данных иметь возможность оспаривать решение.

Оба этих положения требуют предоставления понятных разъяснений. Права на защиту данных связаны с алгоритмической обработкой данных, к таким правам относятся, например: право быть уведомленным об исключительно автоматизированном принятии решений; право на уведомление и доступ к информации о логике, лежащей в основе автоматизированной обработки; право на информацию о значимости и потенциальных последствиях при исключительно автоматизированном принятии решений; право не подвергаться исключительно автоматизированному принятию решений; право оспаривать решение, принятое системой исключительно автоматизированного принятия решений; право на вмешательство человека; право на получение разъяснений.

Аналогично процессам управления данными, набирает силу методика управления самой системой ИИ, в том, насколько дизайн и реализация алгоритмов ИИ согласованы с ценностями и ответственностью организаций и общества.

При появлении компьютерного зрения и, так называемых, «умных» видеокамер и специальных алгоритмов слежения на производстве и в офисах, алгоритмы ИИ начали привлекать для решения вопроса о наложении штрафов на сотрудников [5]. При этом возникают проблемы принятия решений о допуске тех или иных сотрудников к процессам управления и вопросы ответственности [6]. В то же время современные алгоритмы еще не могут приспосабливаться к разным контекстам и решают только те проблемы, которые учтены разработчиками определенным образом.

Заключение

Приведенные выше результаты исследований свидетельствуют о том, что в дискуссиях о месте и роли алгоритмов ИИ не всегда в должной мере оценивается их уязвимость вследствие завышенной оценки их возможностей и надежности. Алгоритмы ИИ часто используются для любых типов автоматизированных систем принятия решений и, к сожалению, в некоторых случаях проявляется тенденция неправомерного применения алгоритмов ИИ в качестве инструмента маркетинга. С другой стороны, такие идеи, как сознательный искус-

ственный интеллект, заменяющий сотрудников практически для любой работы, считается, в настоящее время, универсальным решением проблемы, что далеко неверно и даже неактуально для обсуждения современных алгоритмов ИИ.

Алгоритмы ИИ уже внесли значительный прогресс в таких сферах, как медицина, транспорт, финансы или машинный перевод. Однако, становится важным понять побочные эффекты, которые алгоритмы ИИ неизбежно вызовут после их интеграции в производство продуктов и услуг. Риски для основных прав граждан и организаций серьезны, и алгоритмы ИИ будут иметь значительное влияние на общество. В данной работе показано, что суть этичности алгоритмов ИИ заключается в том, что, принимая критически важные для человека решения, алгоритмы ИИ должны использовать этические императивы, заложенные в алгоритмы обучения.

Рассмотренные выше примеры показывают, что методы анализа и совершенствования систем управления риском на предприятиях и в более общем случае — в предпринимательских экосистемах, активно использующих алгоритмы ИИ, помогают специалистам выявлять и учитывать негативные проявления феномена риска в области этических проблем и обладают достаточными инструментальными средствами для их компенсации.

Список литературы

1. Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта //Философия и общество. – 2018. – №. 2 (87). – С. 84-105. DOI:10.30884/jfio/2018.02.07.

2. Корнина А. Е. Машинное обучение и нейронные сети в бизнесе // Хроноэкономика. – 2018. – №. 2 (10). – С. 111–116.

3. Лобачёва А.С., Соболь О.В. Этика применения искусственного интеллекта в управлении персоналом // E-Management. – 2021. – Т. 4, No 1. – С. 20–28. – DOI:10.26425/2658-3445-2021-4-1-20-28.

4. Becic E., Zych N., Ivarsson J., Vehicle Automation Report HWY18MH010, National Transportation Safety Board – Office of Highway Safety, Tech. Rep., 2019.

5. Hamon R., Junklewitz H., Sanchez I. Robustness and explainability of artificial intelligence – From technical to policy solutions // Luxembourg: Publications Office of the European Union, 2020. – ISBN 978-92-79-14660-5 (online), DOI:10.2760/57493 (online), JRC119336.

6. Kumar N., Kharkwal N., Kohli R., Choudhary S. Ethical aspects and future of artificial intelligence // 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH). – IEEE, 2016. – Pp. 111–114.

7. Varlamov O.O., Chuvikov D.A., Adamova L.E., Petrov M.A., Zabolotskaya I.K., Zhilina T.N. Logical, philosophical and ethical aspects of AI in medicine // International Journal of Machine Learning and Computing. – 2019. – Vol. 9. No. 6. – P. 868.

8. Указ Президента Российской Федерации № 490 от 10 октября 2019 года «О развитии искусственного интеллекта в Российской Федерации». – URL: http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf (дата обращения: 03.10.2022).