Современная Европа, 2024, № 6, с. 156–167

ПРОБЛЕМЫ БЕЗОПАСНОСТИ

УДК 327

РЕГУЛИРОВАНИЕ РИСКОВ, СВЯЗАННЫХ СО ЗЛОНАМЕРЕННЫМ ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В США, ЕС И КИТАЕ[1]

© 2024 БАЗАРКИНА Дарья Юрьевна
Доктор политических наук, ведущий научный сотрудник Отдела исследований европейской интеграции Института Европы РАН и факультета международных отношений Санкт-Петербургского государственного университета 125009, Россия, Москва, Моховая ул., 11-3. ***E-mail:*** *bazarkina-icspsc@yandex.ru.*

© 2024 ПАШЕНЦЕВ Евгений Николаевич
Доктор исторических наук, профессор, главный научный сотрудник Института актуальных международных проблем Дипломатической академии МИД России; профессор факультета международных отношений Санкт-Петербургского государственного университета 107078, Россия, г. Москва, Большой Козловский переулок, 4. ***E-mail:*** *icspsc@mail.ru*

© 2024 МИХАЛЕВИЧ Екатерина Андреевна
Главный специалист Управления организационного развития ПАО «Газпром нефть»; инженер-исследователь факультета международных отношений Санкт-Петербургского государственного университета. 191060, Россия, Санкт-Петербург, ул. Смольного 1/3, подъезд 8. ***E-mail:*** *ekaterina_mikhalevich@mail.ru.*

Поступила в редакцию 08.10.2024
Принята к публикации 05.11.2024

Аннотация. В статье представлен анализ основных механизмов регулирования рисков злонамеренного использования искусственного интеллекта (ЗИИИ) в США, ЕС и КНР. Актуальность проблемы ЗИИИ подтверждена многочисленными данными о применении технологий искусственного интеллекта (ИИ) антисоциальными, преступными акторами, включая террористические. Цель статьи –выявить специфику регулирования рисков злонамеренного использования искусственного интеллекта в США, ЕС и КНР. Установлено, что в США противодействие ЗИИИ еще не выделилось в системные решения на уровне федеральных органов власти, соответствующие меры отдельных штатов и компаний. Речь идет о решениях, которые принимают во

[1] *Работа выполнена при поддержке СПбГУ, шифр проекта 116471555.*

внимание растущие риски ЗИИИ в рамках общего регулирования ИИ и безопасности его использования. В ЕС принят первый в мире закон об искусственном интеллекте, который уделяет незначительное внимание проблемам ЗИИИ. Основными инициаторами предложений по противодействию и регулированию рисков в Евросоюзе выступают силовые структуры, такие как Европол. В КНР регулирование рисков ЗИИИ наиболее централизовано и становится темой стратегических документов и законодательных актов. В статье использован системный подход при рассмотрении различных вариантов угроз ЗИИИ и при формулировке выводов исследования.

Ключевые слова: злонамеренное использование искусственного интеллекта, регулирование искусственного интеллекта, США, Европейский союз, КНР

DOI: 10.31857/S0201708324060147

Проблема злонамеренного (намеренно антисоциального [Pashentsev, 2023: 25]) использования искусственного интеллекта (ЗИИИ) активно обсуждается на уровне государств и интеграционных объединений. Искусственный интеллект (ИИ), будучи технологией двойного назначения, вызывает интерес растущего числа антисоциальных акторов. ИИ уже становится обычным компонентом в арсенале киберпреступников и, скорее всего, получит еще более широкое применение[1], что ставит новые задачи перед правоохранителями и законодателями по всему миру.

Актуальность поставленной проблемы подтверждается ростом количества профильных публикаций в научной и практической сферах. Существует область исследований, в которых рассматриваются проблемы регулирования ИИ [Потемкина, 2019; Базаркина, 2023; Маслова, Сорокова, 2022; Сорокова, 2023; Dixon 2023; Laux, 2023; Antunes, 2024], практика использования и риски ЗИИИ в области кибербезопасности [Hemberg et al, 2020; Malatji, Tolah, 2024; Vassilev et al., 2024], биобезопасности [Cao, 2023, Pauwels, 2023], информационно-психологической безопасности [Doss et al., 2023, Lukacovic, Sellnow-Richmond, 2023; Thomann, 2023], использования ИИ террористами [Bazarkina, 2023, Wells, 2024] и т. п. Угрозам и противодействию ЗИИИ уделено внимание в публикациях Европола[2] и Евроюста[3] в ЕС, государственных органов США[4] и Китая[5].

---

[1] Internet Organised Crime Threat Assessment (IOCTA) 2024. Luxembourg. Publications Office of the European Union. 2024. P. 6. DOI: 10.2813/442713

[2] Facing reality? Law enforcement and the challenge of deepfakes. Luxembourg. Publications Office of the European Union. 2022. 22 p. DOI: 10.2813/158794; ChatGPT – the impact of Large Language Models on Law Enforcement. Luxembourg. Publications Office of the European Union. 2023. 13 p. DOI: 10.2813/255453; Policing in the metaverse: what law enforcement needs to know. Luxembourg. Publications Office of the European Union. 2022. 28 p. DOI: 10.2813/81062

[3] Generative Artificial Intelligence: the impact on intellectual property crimes. The Hague. Eurojust. 2023. 39 p. DOI: 10.2812/1651

[4] Mitigating Artificial Intelligence (AI) Risk: Safety and Security Guidelines for Critical Infrastructure Owners and Operators. US Department of Homeland Security. 2024. 28 p. URL: https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf (дата обращения: 26.08.2024); Increasing Threats of Deepfake Identities. US Department of Homeland Security. URL: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf (дата обращения: 26.08.2024).

[5] Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence. URL: https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html (дата обращения: 26.08.2024); Artificial Intelligence Generated Content (AIGC) White Paper [Rengong

Цель статьи – выявить специфику регулирования рисков ЗИИИ органами законодательной и исполнительной власти в США, ЕС и КНР. Обобщающий термин «регулирование рисков» охватывает различные подходы, включающие «государственное вмешательство в рыночные или социальные процессы для контроля потенциальных неблагоприятных последствий» [Orwat et al., 2024: 10]. Выбор стран и интеграционного объединения продиктован их новаторским опытом в регулировании ИИ.

Регулирование рисков, связанных с ЗИИИ, в США: законодательный бум

Регулирование ИИ в США последние полтора года переживает явный бум. На 1 июня 2024 г. в Конгресс[1] внесены 92 законопроекта в области регулирования искусственного интеллекта. Первые законы, скорее всего, будут приняты после выборов Конгресса и президента в ноябре.

Несколько существующих законов частично регулируют ИИ, например: закон о повторной авторизации Федерального управления гражданской авиации и закон о национальной инициативе в области ИИ на федеральном уровне, калифорнийские правила конфиденциальности (*California Consumer Privacy Act*, *CCPA*) и закон Иллинойса о конфиденциальности биометрической информации[2]. Однако ни в существующих законах, ни в рассматриваемых инициативах нет документа, нацеленного на системное противодействие ЗИИИ. Эта сфера еще не отражена в системных решениях на уровне федеральных органов власти и соответствующих мер отдельных штатов и компаний. Речь идет о решениях, которые принимают во внимание растущие риски ЗИИИ в рамках общего регулирования ИИ и безопасности его использования. За последний год значительную роль в этой области в США сыграли президентские указы.

30 октября 2023 г. президент Дж. Байден издал указ о безопасном, защищенном и заслуживающем доверия ИИ (*Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*). Документ устанавливает «…новые стандарты безопасности и защиты ИИ, обеспечивает конфиденциальность американцев, продвигает справедливость и гражданские права, защищает потребителей…» и обещает «защитить американцев от мошенничества и обмана с применением технологий ИИ…»[3]. Указ ставит ведущих разработчиков искусственного интеллекта под жесткий государственный контроль в соответствии с законом об оборонном производстве (*Defense Production Act*).

zhineng shengcheng neirong (AIGC) baipishu]. URL: http://www.caict.ac.cn/english/research/whitepapers/202211/P020221111501862950279.pdf (дата обращения: 26.08.2024).

[1] 118-й Конгресс США – заседание Конгресса США, действующее с 3 января 2023 г. по 3 января 2025 г.

[2] Babin R. Adapting to AI regulations in the U.S. and Europe: Impacts on CIOs and global enterprises. CIO. 17.06.2024. URL: https://www.cio.com/article/2517656/adapting-to-ai-regulations-in-the-u-s-and-europe-impacts-on-cios-and-global-enterprises.html (дата обращения: 26.08.2024).

[3] Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. White House. 30.10.2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (дата обращения: 26.08.2024).

Компании, которые разрабатывают базовые модели, представляющие серьезную угрозу национальной безопасности, экономической безопасности или общественному здравоохранению, должны уведомлять федеральное правительство при обучении модели и делиться результатами всех тестов безопасности. Очевидная милитаризация вряд ли сможет гармонично сосуществовать со стремлением к продвижению справедливости и гражданских прав в процессах, связанных с разработкой и внедрением ИИ.

В январе 2024 г. администрация Дж. Байдена уведомила об основных действиях в области ИИ после обозначенного указа. Среди них – проект правил, который предлагает обязать американских поставщиков облачных услуг сообщать о предоставлении вычислительных мощностей для обучения ИИ зарубежным разработчикам. «Предложение министерства торговли… потребует от поставщиков облачных услуг уведомлять правительство, когда иностранные клиенты обучают наиболее мощные модели, которые могут быть использованы для вредоносной деятельности»[1]. Неоднозначность формулировки «могут быть использованы для вредоносной деятельности» имеет потенциал лишить иностранных государственных и негосударственных субъектов возможности применить вычислительные мощности США для обучения продвинутых моделей ИИ.

Таким образом, в США два института, которым, согласно опросу Института Гэллапа, не доверяет большинство американцев, – администрация президента и *Big Tech*[2] – собираются контролировать разработку перспективных технологий ИИ, снижая общественный контроль (имеется в виду закон об оборонном производстве) и сужая возможности широкого международного сотрудничества.

В развитие указа о безопасном, защищенном и заслуживающем доверия ИИ Национальный институт стандартов и технологий США (*National Institute of Standards and Technology*, *NIST*) в апреле 2024 г. опубликовал документ «Снижение рисков, связанных с использованием синтетического контента»[3]. В нем рассмотрены существующие стандарты, инструменты и методы, а также потенциал разработки новых научно обоснованных стандартов и методик для аутентификации контента и отслеживания его происхождения; маркировки синтетического контента, например, с помощью водяных знаков; обнаружения синтетического контента; предотвращения создания генеративным ИИ материалов о сексуальном насилии над детьми или интимных изображений без согласия реальных людей и т. п. Меры безопасности в публикации *NIST* не сконцентрированы на возможностях злонамеренного использования синтетического контента, носят исключительно технический характер и не могут считаться руководством по комплексному противодействию ЗИИИ.

В апреле 2024 г. министерство внутренней безопасности США учредило Совет по охране и безопасности ИИ из 35 человек, который будет предоставлять рекомендации по повышению безопасности и устойчивости критически важной инфраструктуры при использовании технологий искусственного интеллекта. В качестве противовеса участию руководителей *Big Tech* в Совете министерство также выбрало несколько извест-

[1] Там же.

[2] Saad L. Historically Low Faith in U.S. Institutions Continues. Gallup. 06.07.2023. URL: https://news.gallup.com/poll/508169/historically-low-faith-institutions-continues.aspx (дата обращения: 29.02.2024).

[3] Reducing Risks Posed by Synthetic Content. An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4. National Institute of Standards and Technology, US Department of Commerce. 04.2024. URL: https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf (дата обращения: 26.08.2024).

ных представителей организаций по защите гражданских прав и научных кругов. Пока не ясно, будет ли достаточно голосов последних для того, чтобы сдержать связку высшего чиновничества и *Big Tech*.

В целом различные министерства и ведомства США все активнее подключаются к процессу регулирования ИИ. Этот процесс усилился не только в рамках выполнения президентских указов, но и по причине роста рисков и угроз использования ИИ, озабоченности широких слоев населения и крупного бизнеса этими угрозами.

Подход ЕС к рискам, связанным с ЗИИИ: Закон ЕС об ИИ и публикации Европола

21 мая 2024 г. Совет ЕС официально утвердил проект закона об ИИ[1]. 1 августа 2024 г. вступили в силу его первые положения. Закон введет ряд новых обязательств для компаний, предоставляющих, распространяющих, импортирующих и использующих системы ИИ в Евросоюзе. Их невыполнение грозит штрафами в размере до 35 млн евро, или 7% от общего мирового годового оборота.

В законе ЕС об ИИ слово «злонамеренный» (*malicious*) встречается два раза. В статье 50 указано, что системы искусственного интеллекта с высоким уровнем риска должны быть устойчивы к собственным ограничениям и злонамеренным действиям. В статье 51 заявлено, что кибербезопасность «играет решающую роль в обеспечении устойчивости систем ИИ к попыткам … поставить под угрозу их свойства безопасности со стороны злонамеренных третьих лиц». Подчеркивается, что кибератаки на системы искусственного интеллекта могут использовать наборы данных для обучения (например, при отравлении данных), уже обученные модели ИИ, а также «уязвимости в цифровых активах системы ИИ или базовой инфраструктуре ИКТ»[2]. В приложении I к закону перечислены восемь областей, в которых используются различные технологии ИИ высокого риска, подлежащие регулированию[3]. Самый обширный список представлен для правоохранительной отрасли. Он включает системы ИИ, предназначенные для оценки риска совершения первичного или повторного правонарушения физическими лицами или риска стать потенциальной жертвой преступлений; использования правоохранительными органами в качестве полиграфов или для определения эмоционального состояния физического лица; обнаружения дипфейков; прогнозирования преступления на основе профилирования физических лиц[4] и т. п. Таким образом, прежде всего контролируется легальное и незлонамеренное использование ИИ.

Большое внимание ЗИИИ уделяют силовые ведомства. Европол опубликовал ряд отчетов, разработанных инновационным центром ЕС по внутренней безопасности, который функционирует в рамках агентства. Научно-исследовательские проекты центра включают, в частности, ИИ, биометрию, шифрование, разведку по открытым источникам, безопасность связи, виртуальную и дополненную реальность. Один из главных проектов –

[1] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (дата обращения: 10.08.2024).

[2] Там же.

[3] Там же.

[4] Там же.

разработка принципов подотчетности ИИ в сфере внутренней безопасности[1] – стартовал в 2021 г. Его цель – разработать решения для оценки, анализа и обеспечения ответственности за использование ИИ специалистами по внутренней безопасности в соответствии с ценностями и основными правами Евросоюза.

По оценкам Европола, в ЕС растет угроза злонамеренного использования дипфейков. Например, правоохранительные органы Испании в 2023 г. рассмотрели первые случаи манипуляции изображениями с помощью ИИ, связанные с кибербуллингом. Подозреваемые (несовершеннолетние, некоторые моложе 14 лет) сфотографировали не менее 22 молодых женщин и преобразовали фото с помощью приложения на базе ИИ в изображения сексуального характера. Затем изображения широко распространялись в социальных сетях и приложениях для общения. Жертвы получили значительный психологический ущерб[2]. Этот случай свидетельствует о растущих возможностях использования ИИ преступниками, что в дальнейшем может вывести дипфейки из категории систем ИИ ограниченного риска.

Инновационный центр ЕС по внутренней безопасности ставит актуальные проблемы в области противодействия ЗИИИ. По мнению правоохранителей, «изменение материалов в социальных сетях о таких мероприятиях, как митинги, может привести к тому, что полиция начнет действовать … не в том месте»[3], например, преследовать не того человека, если фейковая версия подозреваемого становится вирусной в социальных сетях. Используя дипфейки, злоумышленники могут изображать сотрудников полиции, совершающих правонарушения. «В эпоху, когда недоверие к властям растет»[4], дипфейки могут использоваться для негативного воздействия на общественное мнение. По признанию Европола, влияние таких изображений и видеозаписей нельзя недооценивать, особенно когда это сочетается с доксингом (раскрытием личности) офицеров.

Полицейское агентство ЕС с вниманием относится к использованию больших языковых моделей злоумышленниками. Европол отмечает, что способность *ChatGPT* составлять высоко аутентичные тексты на основе пользовательского запроса делает его чрезвычайно полезным инструментом фишинга. С помощью больших языковых моделей манипулятивные сообщения создаются быстрее и выглядят достовернее. Возможности *ChatGPT* растут также «в области терроризма, пропаганды и дезинформации»[5]. Модель может использоваться в сборе информации для террористической деятельности, например финансирования терроризма.

Европол обеспокоен тем, что террористы будут использовать новые технологические возможности в случае создания метавселенной. По мере того как становится доступным все более точный и полный цифровой двойник реальности, он может предоставлять информацию о целях терактов. В будущем это может позволить проводить военную разведку и планирование в пределах метавселенной. С другой стороны, метавселенная может позволить пользователям создать виртуальный мир таким, каким они его себе представляют, например, виртуальный халифат или «государство превосходства белой расы». Так, нацистские газовые камеры уже были созданы во вселенной

[1] Accountability Principles for Artificial Intelligence. URL: https://www.ap4ai.eu/ (дата обращения: 26.08.2024).
[2] Internet Organised Crime Threat Assessment (IOCTA). P. 26.
[3] Facing reality? Law enforcement and the challenge of deepfakes. P. 14.
[4] Там же.
[5] ChatGPT – the impact of Large Language Models on Law Enforcement. P. 7–8.

«Роблокс» (*Roblox*)[1]. Более того, пространства метавселенной обеспечат идеальную среду для вербовки террористов в других виртуальных мирах и физическом мире[2]. Силовые структуры ЕС обеспокоены рисками ЗИИИ больше, чем высшие правительственные органы.

Технологическое сообщество и общество в целом недовольно законодательством ЕС. Как заявил Макс фон Тун, европейский директор Института открытых рынков, «закон об ИИ не способен устранить угрозу номер один, которую в настоящее время представляет ИИ: его роль в увеличении и закреплении чрезвычайной власти, которую несколько доминирующих технологических компаний уже имеют в нашей личной жизни, нашей экономике и наших демократиях». ЗИИИ в информационной среде вызывают растущее опасение. «Системы искусственного интеллекта, используемые в информационном пространстве, должны быть классифицированы как системы высокого риска, … чего прямо не предусмотрено в принятом Законе ЕС об ИИ», – отметила Катарина Цюгель, менеджер по политике Форума «Информация и демократия»[3].

Таким образом, ЕС делает акцент прежде всего на регулировании поведения разработчиков ИИ, а противодействие ЗИИИ выражено в рекомендациях полицейских органов. В будущем такой подход может привести к расширению существующего законодательства в области искусственного интеллекта по мере того, как будет расти число злоупотреблений.

Регулирование рисков, связанных с ИИ, в КНР

Китайское правительство закрепляет общие политические положения по развитию и регулированию технологий искусственного интеллекта в стратегических документах. В 2022 г. Китайская академия информационных и коммуникационных технологий при министерстве промышленности и информационных технологий опубликовала Белую книгу о контенте, создаваемом ИИ. Целый раздел посвящен проблемам, с которыми сталкивается развитие этой сферы, в т. ч. действиям злоумышленников[4].

Угроза распространения заведомо ложного контента, созданного ИИ, диктует необходимость разработки цифровых продуктов для противодействия. На этом направлении активно работают технологические гиганты КНР, спонсируемые правительством. К примеру, в начале 2023 г. запущена бета-версия первого в Китае инструмента обнаружения ложного контента. Технология *AIGC-X* способна обнаруживать фейковые новости и спам, генерируемые ИИ, с точностью до 90%[5] и имеет широкие перспективы применения в области безопасности контента, защиты авторских прав и интеллектуальной собственности, а также предотвращения фишинговых атак и распространения

[1] *Roblox* – онлайн-платформа, которая позволяет пользователям программировать собственные игры и играть в созданные другими пользователями.

[2] Policing in the metaverse: what law enforcement needs to know. P. 19.

[3] EU AI Act reaction: Tech experts say the world's first AI law is “historic” but “bittersweet”. Euronews. 16.03.2024. URL: https://www.euronews.com/next/2024/03/16/eu-ai-act-reaction-tech-experts-say-the-worlds-first-ai-law-is-historic-but-bittersweet (дата обращения: 26.08.2024).

[4] Artificial Intelligence Generated Content (AIGC) White Paper. URL: http://www.caict.ac.cn/english/research/whitepapers/202211/P020221111501862950279.pdf (дата обращения: 26.08.2024).

[5] About AIGC-X [Guanyu AIGC-X]. 2023. URL: http://ai.sklccc.com/AIGC-X/#/ (дата обращения: 26.08.2024).

фейковых новостей[1]. Ближайшие цели разработки – настройка обнаружения контента на английском языке (в бета-версии доступен только китайский), улучшение возможностей обнаружения аудио- и видеоконтента, генерируемого ИИ[2].

Многие крупные технологические компании КНР принимают активное участие в развитии безопасных технологий ИИ. «Дангхонг Текнолоджи» (*Danghong Technology*) в ответ на популяризацию *ChatGPT* объявила о разработке приложений, позволяющих генерировать контент на основе ИИ, а также программ, идентифицирующих такой контент. «Бохуи Текнолоджи» (*Bohui Technology*), компания по обеспечению безопасности СМИ, разрабатывает интеллектуальные решения для создания опыта обучения на основе ИИ в образовательных учреждениях, который повысит общую осведомленность и углубит навыки распознавания контента, созданного с помощью искусственного интеллекта.

По мнению китайского правительства, приоритетными направлениями противодействия рискам ЗИИИ должны стать разработка национальной, а в последующем и международной нормативно-правовой базы, регулирующей сферу ИИ; создание системы социальной этики путем обучения граждан информационной грамотности, а также осуществление общественного контроля посредством создания системы социального доверия. В марте 2022 г. вступило в силу положение об администрировании рекомендаций по алгоритмам информационных услуг Интернета. Статья 9 обязует поставщиков услуг усиливать управление информационной безопасностью, создавать и улучшать базы данных для выявления незаконной и вредной информации, а также улучшать процедуры входа в базы данных. Положение предупреждает об ответственности за использование алгоритмов интернет-рекомендаций в злонамеренных целях (штрафы на сумму от 10 до 100 тыс. юаней)[3].

Большое значение для предотвращения ЗИИИ имеет положение об администрировании информационных интернет-сервисов глубокого синтеза, вступившее в силу 10 января 2023 г. Под термином «технология глубокого синтеза» в документе понимаются технологии, которые на основе алгоритма синтеза, представленного машинным обучением и возможностями виртуальной реальности, способны создавать уникальные продукты, в т. ч. создавать и редактировать текстовый, голосовой и неголосовой (звуки, музыка) контент, генерировать, заменять лица, манипулировать последними и т. п. Поставщики и пользователи услуг глубокого синтеза не должны использовать такие технологии для производства, копирования, публикации или распространения ложной новостной информации[4].

Китай подчеркивает важность международного сотрудничества для противостояния ЗИИИ. В конце 2022 г. МИД КНР опубликовал позиционный документ по усилению этического управления ИИ, в котором провозглашается обязательство поощрять иссле-

[1] AI generated content can be detected [AI shengcheng neirong keyi jiance]. 2023. URL: https://www.163.com/dy/article/HUOPFFAG0514R9M0.html (дата обращения: 26.08.2024).

[2] About AIGC-X...

[3] Provisions on the Administration of Internet Information Service Algorithm Recommendations [Hulianwang xinxi fuwu suanfa tuijian guanli guiding]. 2022. URL: http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm (дата обращения: 26.08.2024).

[4] The State Internet Information Office and other three departments issued the “Regulations on the Administration of Deep Synthesis of Internet Information Services”. URL: http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm (дата обращения: 26.08.2024).

дования и разработки и продвигать международное сотрудничество. В документе отмечается, что развитие технологий ИИ должно быть «ориентировано на людей» и следовать принципу «ИИ во благо»[1].

Помимо конкретных правовых и технических решений по нейтрализации угроз ЗИИИ, общим условием успешной нейтрализации этих угроз выступают меры по обеспечению социально ориентированного развития Китая. Они объективно сужают возможности внутренних и внешних антисоциальных акторов использовать ИИ в злонамеренных целях. Кроме того, Пекин стремится избежать чрезмерного контроля над поставщиками услуг с технологиями ИИ и потребителями, так как это может помешать Китаю стать мировым лидером в области искусственного интеллекта, а также создать серьезные трудности для социально-экономического развития страны и укрепления национальной безопасности.

Заключение

В США, ЕС и КНР растет актуальность проблем ЗИИИ, что выражено в законодательной практике и рекомендациях органов безопасности. В Соединенных Штатах регулирование ИИ переживает бум. В Евросоюзе принят первый в мире закон об ИИ, в котором уделено внимание ЗИИИ. В КНР тема ЗИИИ активно разрабатывается в стратегических документах. В поле зрения правительств попадают разные угрозы – от кибербуллинга до манипуляций с автоматизированными рекомендациями онлайн-контента. Опасения вызывает злонамеренное использование дипфейков и манипуляции в метавселенных, в т. ч. со стороны террористических организаций.

Регулирование рисков, связанных с ЗИИИ, в трех исследованных юрисдикциях находится на начальном этапе формирования. Российским профильным структурам может быть полезен опыт социально-ориентированного регулирования КНР, активных дискуссий в США и попыток ЕС предложить Закон об ИИ как образец для других стран. Все это нуждается в критическом осмыслении, так как даже в КНР с централизованной системой государственного управления еще не решена проблема монополий технологических гигантов. В США, Китае и Евросоюзе сохраняются угрозы распространения террористической пропаганды, которые больше связаны с социальными проблемами, чем с распространением ИИ. Проблемы, вызванные ЗИИИ, требуют комплексных решений с участием как разработчиков ИИ и правоохранителей, так и широких слоев гражданского общества.

СПИСОК ЛИТЕРАТУРЫ

Базаркина Д.Ю. (2023) Цифровое десятилетие Европейского союза: цели и промежуточные результаты. *Аналитические записки ИЕ РАН*. Выпуск 4. С. 89–98. DOI: http://doi.org/10.15211/analytics43420238998

Потемкина О.Ю. (2019) Лучше, чем люди? Политика ЕС в области искусственного интеллекта. *Научно-аналитический вестник ИЕ РАН*. № 5. С. 16–21. DOI: http://dx.doi.org/10.15211/vestnikieran520191621

Маслова Е.А., Сорокова Е.Д. (2022) Диалектика этики и права в регулировании технологии искусственного интеллекта: опыт ЕС. *Современная Европа*. № 5(112). С. 19–33. DOI: 10.31857/S0201708322050023

[1] Position Paper of the People's Republic of China…

Сорокова Е.Д. (2023) *Реагирование Европейского союза на глобальные технологические риски: практика и эффективность*. Дис. … канд. полит. наук. МГИМО, Москва. 280 с.

Antunes H.S. (2024) European AI Regulation Perspectives and Trends. *Legal Aspects of Autonomous Systems. ICASL 2022. Data Science, Machine Intelligence, and Law.* Ed. by D. Moura Vicente, R. Soares Pereira, A. Alves Leal. Vol. 4. Springer, Cham, Switzerland. P. 53–65. DOI: https://doi.org/10.1007/978-3-031-47946-5_4

Bazarkina D. (2023) Current and Future Threats of the Malicious Use of Artificial Intelligence by Terrorists: Psychological Aspects. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Ed. by E. Pashentsev E. Palgrave Macmillan, Cham, Switzerland. P. 251–272. DOI: https://doi.org/10.1007/978-3-031-22552-9_10

Cao L. (2023) AI and data science for smart emergency, crisis and disaster resilience. *International Journal of Data Science and Analytics.* No. 15. P. 231–246. DOI: https://doi.org/10.1007/s41060-023-00393-w

Dixon R.B.L. (2023) A principled governance for emerging AI regimes: lessons from China, the European Union, and the United States. *AI Ethics.* No. 3. P. 793–810. DOI: https://doi.org/10.1007/s43681-022-00205-0

Doss C., Mondschein J., Shu D. Wolfson T., Kopecky D., Fitton-Kane V. A., Bush L., Tucker C. (2023) Deepfakes and scientific knowledge dissemination. *Scientific Reports.* No. 13. 13429. DOI: https://doi.org/10.1038/s41598-023-39944-3

Hemberg E., Zhang L., O'Reilly U.M. (2020) Exploring Adversarial Artificial Intelligence for Autonomous Adaptive Cyber Defense. *Adaptive Autonomous Secure Cyber Systems*. Ed. by S. Jajodia, G. Cybenko, V. Subrahmanian, V. Swarup, C. Wang, M. Wellman. Springer, Cham, Switzerland. P. 41–61. DOI: https://doi.org/10.1007/978-3-030-33432-1_3

Laux J. (2023) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & Society*. DOI: https://doi.org/10.1007/s00146-023-01777-z

Lukacovic M.N., Sellnow-Richmond D.D. (2023) The COVID-19 Pandemic and the Rise of Malicious Use of AI Threats to National and International Psychological Security. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Ed. by E. Pashentsev. Palgrave Macmillan, Cham, Switzerland. P. 175–201. DOI: https://doi.org/10.1007/978-3-031-22552-9_7

Malatji M., Tolah A. (2024) Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI Ethics*. DOI: https://doi.org/10.1007/s43681-024-00427-4

Orwat C., Bareis J., Folberth A., Jahnel J., Wadephul C. (2024) Normative Challenges of Risk Regulation of Artificial Intelligence. *Nanoethics.* Vol. 18. Article 11. DOI: https://doi.org/10.1007/s11569-024-00454-9

Pashentsev E. (2023) General Content and Possible Threat Classifications of the Malicious Use of Artificial Intelligence to Psychological Security. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Ed. by E. Pashentsev. Palgrave Macmillan, Cham, Switzerland. P. 23–46. DOI: https://doi.org/10.1007/978-3-031-22552-9_2

Pauwels E. (2023) How to Protect Biotechnology and Biosecurity from Adversarial AI Attacks? A Global Governance Perspective. *Cyberbiosecurity*. Ed. by D. Greenbaum. Springer, Cham, Switzerland. P. 173–184. DOI: https://doi.org/10.1007/978-3-031-26034-6_11

Thomann P.E. (2023) Geopolitical Competition and the Challenges for the European Union of Countering the Malicious Use of Artificial Intelligence. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Ed. by E. Pashentsev. Palgrave Macmillan, Cham, Switzerland. P. 453–486. DOI: https://doi.org/10.1007/978-3-031-22552-9_17

Vassilev A., Oprea A., Fordyce A., Anderson H. (2024) *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.* NIST Artifcial Intelligence (AI) Report, NIST AI 100-2e2023. National Institute of Standards and Technology, Gaithersburg, USA. 99 p. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023

Wells D. (2024) *The Next Paradigm-Shattering Threat? Right-Sizing the Potential Impacts of Generative AI on Terrorism***.** Middle East Institute, Washington, USA. 16 p.

Regulating the Risks Associated with Malicious Use of Artificial Intelligence in the US, EU and China[1]

D.Yu. Bazarkina
Doctor of Sciences (Politics), Leading Researcher, Department of European Integration Research, Institute of Europe, Russian Academy of Sciences; School of International Relations, Saint Petersburg State University. 11-3, Mokhovaya Street, Moscow, Russia, 125009.
E-mail: *bazarkina-icspsc@yandex.ru.*

E.N. Pashentsev
Doctor of Sciences (History), Professor, Chief Researcher, Institute of Contemporary International Studies of the Diplomatic Academy of the Ministry of Foreign Affairs of Russia; Professor, Faculty of International Relations, St. Petersburg State University.
4, Bolshoi Kozlovsky Lane, Moscow, Russia, 107078; ***E-mail:*** *icspsc@mail.ru.*

E.A. Mikhalevich
Chief specialist of the Organizational Development Department of Gazprom Neft PJSC; Research engineer at the School of International Relations, Saint Petersburg State University; 1/3 Smolny Street, entrance 8, Saint Petersburg, Russia, 191060
E-mail: *ekaterina_mikhalevich@mail.ru*

Abstract. The article presents an analysis of the main mechanisms for regulating risks caused by the malicious use of artificial intelligence (MUAI) in the USA, the EU and China. The relevance of the MUAI problem is proven by numerous data on the use of artificial intelligence (AI) technologies by antisocial actors. The authors set a goal – to identify the specifics of regulating MUAI risks in the USA, EU and China – due to the innovative experience of these three jurisdictions. It was found that in the USA counteraction to MUAI has not yet been shaped into systemic decisions at the level of federal authorities. It is more about decisions that take into account the growing risks of MUAI within the framework of general regulation of AI and the safety of its use. The EU has adopted the world's first Law on AI, which however pays little attention to the MUAI issues, and the main initiators of proposals to counter and regulate risks are law enforcement agencies, such as Europol. In China, MUAI risk regulation is most centralized and is becoming the subject of strategic documents and legislative acts. The authors use a systemic approach when considering various options for MUAI threats and formulating the research conclusions.

Keywords: malicious use of artificial intelligence, regulation of artificial intelligence, USA, European Union, PRC

DOI: 10.31857/S0201708324060147

REFERENCES

Antunes H.S. (2024) European AI Regulation Perspectives and Trends, in Moura Vicente D., Soares Pereira R., Alves Leal A. (ed.) *Legal Aspects of Autonomous Systems. ICASL 2022. Data Science, Machine Intelligence, and Law*, 4, Springer, Cham, Switzerland, pp. 53–65. DOI: https://doi.org/10.1007/978-3-031-47946-5_4

Bazarkina D. (2023) Current and Future Threats of the Malicious Use of Artificial Intelligence by Terrorists: Psychological Aspects, in Pashentsev E. (ed.) *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, Palgrave Macmillan, Cham, Switzerland, pp. 251–272. DOI: https://doi.org/10.1007/978-3-031-22552-9_10

[1] *The authors acknowledge Saint-Petersburg State University for a research project 116471555.*

Bazarkina D. (2023) Tsifrovoye desyatiletiye Yevropeyskogo soyuza: tseli i promezhutochnyye rezul'taty [The European Union's Digital Decade: Objectives and Interim Results], *Analytical notes of IE RAS,* 4, pp. 89–98. DOI: http://doi.org/10.15211/analytics43420238998 (In Russian).

Cao L. (2023) AI and data science for smart emergency, crisis and disaster resilience, *International Journal of Data Science and Analytics,* 15, pp. 231–246. DOI: https://doi.org/10.1007/s41060-023-00393-w

Dixon R.B.L. (2023) A principled governance for emerging AI regimes: lessons from China, the European Union, and the United States, *AI Ethics*, 3, pp. 793–810. DOI: https://doi.org/10.1007/s43681-022-00205-0

Doss C., Mondschein J., Shu D. Wolfson T., Kopecky D., Fitton-Kane V. A., Bush L., Tucker C. (2023) Deepfakes and scientific knowledge dissemination, *Scientific Reports*, 13, 13429. DOI: https://doi.org/10.1038/s41598-023-39944-3

Hemberg E., Zhang L., O'Reilly U.M. (2020) Exploring Adversarial Artificial Intelligence for Autonomous Adaptive Cyber Defense, in Jajodia S., Cybenko G., Subrahmanian V., Swarup V., Wang C., Wellman M. (ed.) *Adaptive Autonomous Secure Cyber Systems*, Springer, Cham, Switzerland, pp. 41–61. DOI: https://doi.org/10.1007/978-3-030-33432-1_3

Laux J. (2023) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act, *AI & Society*, DOI: https://doi.org/10.1007/s00146-023-01777-z

Lukacovic M.N., Sellnow-Richmond D.D. (2023) The COVID-19 Pandemic and the Rise of Malicious Use of AI Threats to National and International Psychological Security, in Pashentsev E. (ed.) *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, Palgrave Macmillan, Cham, Switzerland, pp. 175–201. DOI: https://doi.org/10.1007/978-3-031-22552-9_7

Malatji M., Tolah A. (2024) Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI, *AI Ethics*. DOI: https://doi.org/10.1007/s43681-024-00427-4

Maslova E., Sorokova E. (2022) Dialektika etiki i prava v regulirovanii tekhnologii iskusstvennogo intellekta: opyt ES [The Dialectics of Ethics and Law in the Regulation of Artificial Intelligence: Case of the EU], *Contemporary Europe,* 5(112), pp. 19–33. DOI: https://doi.org/10.31857/S0201708322050023 (In Russian).

Orwat C., Bareis J., Folberth A., Jahnel J., Wadephul C. (2024) Normative Challenges of Risk Regulation of Artificial Intelligence, *Nanoethics*, 18(11). DOI: https://doi.org/10.1007/s11569-024-00454-9

Pashentsev E. (2023) General Content and Possible Threat Classifications of the Malicious Use of Artificial Intelligence to Psychological Security, in Pashentsev E. (ed.) *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, Palgrave Macmillan, Cham, Switzerland, pp. 23–46 DOI: https://doi.org/10.1007/978-3-031-22552-9_2

Pauwels E. (2023) How to Protect Biotechnology and Biosecurity from Adversarial AI Attacks? A Global Governance Perspective, in Greenbaum D. (ed.) *Cyberbiosecurity*, Springer, Cham, Switzerland, pp. 173–184. DOI: https://doi.org/10.1007/978-3-031-26034-6_11

Potemkina O. (2019) Luchshe, chem lyudi? Politika ES v oblasti iskusstvennogo intellekta [Better than Humans? EU Artificial Intelligence Policy], *Nauchno-analiticheskij vestnik IE RAN*, 5, pp. 16–21. DOI: http://dx.doi.org/10.15211/vestnikieran520191621 (In Russian).

Sorokova E. (2023) *Reagirovaniye Yevropeyskogo soyuza na global'nyye tekhnologicheskiye riski: praktika i effektivnost'* [The European Union's Response to Global Technological Risks: Practice and Efficiency], PhD Thesis, MGIMO, Moscow, Russia. (In Russian).

Thomann P.E. (2023) Geopolitical Competition and the Challenges for the European Union of Countering the Malicious Use of Artificial Intelligence, in Pashentsev E. (ed.) *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, Palgrave Macmillan, Cham, Switzerland, pp. 453–486. DOI: https://doi.org/10.1007/978-3-031-22552-9_17

Vassilev A., Oprea A., Fordyce A., Anderson H. (2024) *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST Artifcial Intelligence (AI) Report, NIST AI 100-2e2023, National Institute of Standards and Technology, Gaithersburg, USA. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023

Wells D. (2024) *The Next Paradigm-Shattering Threat? Right-Sizing the Potential Impacts of Generative AI on Terrorism*, Middle East Institute, Washington, USA.