УДК 004.942
EDN: CYVYLQ

ПРИНЯТИЕ ОТВЕТСТВЕННЫХ РЕШЕНИЙ И РИСКИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

SPIN-код: 7198-5521
AuthorID: 415365
ORCID: 0000-0002-7495-1090

ПРОКОФЬЕВ Олег Владимирович, кандидат технических наук,
доцент кафедры «Информационные технологии и системы»
Пензенский государственный технологический университет
(440039, Россия, Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11, e-mail Prokof_ow@mail.ru)

Аннотация. За последние два десятилетия область искусственного интеллекта (ИИ) развивалась все более быстрыми темпами. Системы, использующие интеллектуальные технологии, затронули многие аспекты жизни граждан страны. Неудивительно, что ИИ также предлагает большие перспективы в чувствительных областях с точки зрения жизни и здоровья человека. Растущее число роботизированных транспортных средств, устройств клинического назначения, систем для работы в опасных условиях и других интеллектуальных устройств приводит к необходимости частого принятия ответственных решений, риски от которых увеличиваются по мере роста автономности систем и устройств. Интеллектуальные защитные системы все чаще обнаруживают, анализируют и реагируют на изменения среды быстрее и эффективнее, чем люди-операторы. А системы анализа больших данных и поддержки принятия решений дают возможность усвоить большие объемы информации, которые не смогла бы усвоить ни одна группа аналитиков, какой бы многочисленной она ни была, и помочь лицам, принимающим ответственные решения, быстрее выбирать более эффективные направления действий. Одновременно в странах мира происходит разработка этических кодексов, правил и норм ИИ, проводится регулирование правовой оценки ответственности специалистов по машинному обучению. Активное внедрение систем в ИИ в сфере вооружений в разных странах мира представляет направление особо высокого риска, если не будет уделено должное внимание надёжности, безопасности и гуманитарным последствиям. Исследования в этой области пока не опираются на большой опыт применения, чаще они основываются на мнение экспертов – разработчиков и исследователей автономных систем с ИИ. Статья посвящена оценкам рисков, сопутствующих применению систем ответственного назначения на базе ИИ.

Ключевые слова: риски искусственного интеллекта, принятие ответственных решений, автономные системы, этические нормы искусственного интеллекта.

RESPONSIBLE DECISION-MAKING AND THE RISKS OF ARTIFICIAL INTELLIGENCE

PROKOFIEV Oleg Vladimirovich, candidate of technical sciences,
associate professor of the Department of Information Technologies and Systems
Penza State Technological University
(440039, Russia, Penza, BaydukovProyezd / Gagarin Street, 1a/11, e-mail Prokof_ow@mail.ru)

Abstract. Over the past two decades, the field of artificial intelligence (AI) has developed at an increasingly rapid pace. Systems using intelligent technologies have affected many aspects of the life of the country's citizens. Not surprisingly, AI also offers great promise in sensitive areas in terms of human life and health. The growing number of robotic vehicles, clinical devices, hazardous environment systems, and other intelligent devices is driving the need for frequent critical decision-making, the risks of which increase as the autonomy of systems and devices increases. Intelligent security systems are increasingly detecting, analyzing and responding to environmental changes faster and more efficiently than human operators. And big data analytics and decision support systems make it possible to assimilate large amounts of information that no single group of analysts, no matter how large, could assimilate, and help decision makers to choose more effective courses of action faster. At the same time, ethical codes, rules and norms of AI are being developed in the countries of the world, and the legal assessment of the responsibility of machine learning specialists is being regulated. The active implementation of systems in AI in the field of weapons around the world represents a particularly high-risk direction if due attention is not paid to reliability, security and humanitarian consequences. Research in this area is not yet based on extensive application experience, more often they are based on the opinion of experts - developers and researchers of autonomous systems with AI. The article is devoted to risk assessments associated with the use of responsible AI-based systems.

Keywords: risks of artificial intelligence, responsible decision-making, autonomous systems, ethical standards of artificial intelligence.

Для цитирования: *Прокофьев О.В. Принятие ответственных решений и риски искусственного интеллекта/ О.В. Прокофьев// XXI век: итоги прошлого и проблемы настоящего плюс. – 2023. – Т. 12. – № 3(63). – С. 26-31. – EDN: CYVYLQ.*

Введение. Важное различие, которое следует сделать при обсуждении приложений ИИ, заключается в том, действительно ли система автономной или просто автоматизированной. Когда люди хотят, чтобы задача была выполнена, они делают это сами или делегируют ее другому объекту, которым может быть человек или машина. При делегировании они отказываются от некоторого контроля над тем, как это делается, и лицо, выполняющее задачу, получает некоторую степень автономии. Если же задача идеально сформулирована с набором определенных и известных правил, то объект, выполняющий ее, имеет «низкую автономию» и её характеризуют как «автоматизированную».

Кроме того, люди, работающие в области ИИ, часто различают то, что они называют автономия в состоянии покоя и автономность в динамическом режиме, когда взаимодействуют с физическим миром. Примеры автономии в динамике, вызвавшие закономерную обеспокоенность, включают автономные системы вооружения. Поскольку ИИ охватывает так много видов систем и уровней автономии, полезно классифицировать эти технологии с помощью графической таксономии, иллюстрирующей отношения между ними (рис. 1).

Ранние подходы к ИИ включали разработку автоматизированных систем, способных выполнять задачи по сценарию в соответствии с наборами заданных правил. Такие подходы все еще используются до некоторой степени, но за последние пару десятилетий были разработаны более сложные системы, способные к машинному обучению (МО). Эти системы могут постепенно повышать свою производительность, распознавая закономерности в больших объемах данных и предпринимая корректирующие действия для улучшения их работы. Еще более сложный класс систем машинного обучения демонстрирует глубокое обучение. Они используют многослойные искусственные нейронные сети для распознавания шаблонов в представлениях данных, недавние прорывы в области глубокого обучения с использованием глубоких нейронных сетей позволили добиться значительных успехов в системах компьютерного зрения и распознавания изображений.

Поэтому целью проделанной работы является анализ оценок рисков, связанных с автономным использованием систем на базе ИИ в областях ответственного применения. В качестве области применения выбрана сфера автономных вооружений как наиболее чувствительная с точки зрения сохранения жизни и здоровья, гуманитарных ценностей и этических правил.

Рисунок 1 – Таксономия технологий искусственного интеллекта

Методология. В качестве источника информации были использованы результаты экспертных оценок Института мира [1-11] и результаты опросов специалистов, интервью, проведённые американской корпорацией RAND [12]. В последнем случае использованы знания 29 экспертов в области ИИ, в числе которых среди опрошенных были военные разработчики и операторы, лидеры мнений в бизнесе и академических кругах, бывшие чиновники Министерства обороны США, а также отставные генералы военно-воздушных сил и сухопутных войск. Опросы были составлены таким образом, чтобы выяснить, что думают эксперты о будущем военного применения ИИ, может ли ИИ изменить характер войны и как скоро это произойдёт, а также какие опасения они формулируют, если они имеются. Важно выяснить, существуют или предполагаются преимущества использования ИИ, без которых описание ситуации риска не совсем полно и объективно. Предварительный анализ, уточняющий

построение вопросов и формулировки терминов, привёл к систематизированному описанию рисков как таковых. Поэтому в процедуру сбора и обработки данных было введено уточнение относительно конкретной предметной области, определяющей сущность преимуществ и рисков автономной системы. Классификация рисков по области появления событий, связанных с ущербом в отношении человеческой жизни и здоровья, а также с потерями в области задач военного применения представлена на рисунке 2. Перечень разновидностей рисков подтверждает озабоченность авторов открытого письма учёных, специалистов в области искусственного интеллекта [13], призывающих к тому, чтобы современные мощные системы ИИ стали более точными, безопасными, интерпретируемыми, прозрачными, надежными, согласованными, заслуживающими доверия и лояльными.

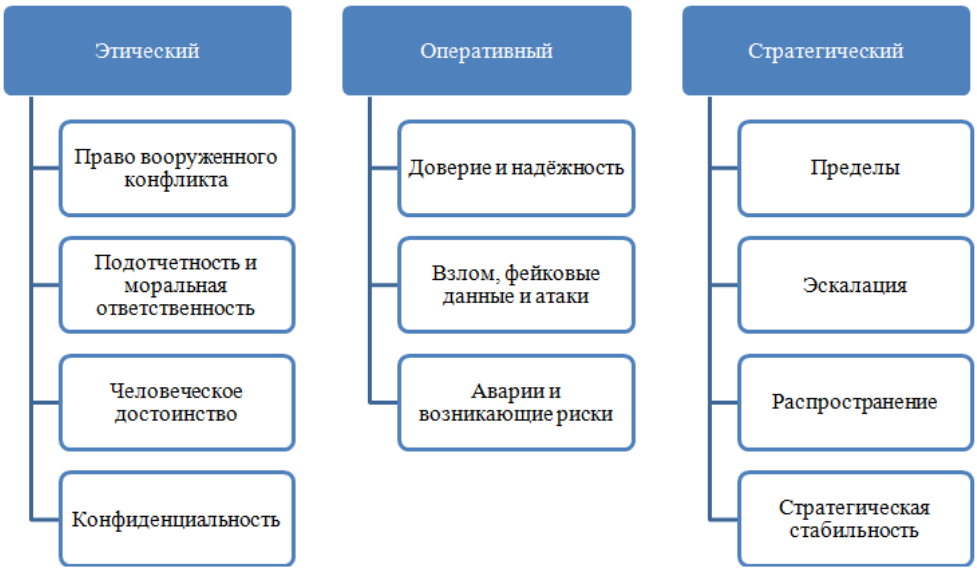

Рисунок 2 – Классификация рисков искусственного интеллекта

Результаты. Потенциальные преимущества использования ИИ, названные респондентами *RAND*, изображены на рисунке 3. Похожие ответы были сгруппированы. Наиболее часто упоминаемое преимущество представляет собой высокую скорость принятия решений. С другой стороны, преобладает время, необходимое для перемещения оборудования или людей, а иногда и время, в течение которого боеприпасы перемещаются к целям. Еще один риск заключается в том, что если скорость сделать приоритетным атрибутом при выборе между конкурирующими автономными системами вооружения для разработки, это может быть достигнуто за счёт безопасности и надежности.

Использование больших данных в непосредственном виде трудно доступно человеку, но машины и ИИ, как правило, работают лучше, чем больше данных им доступно. Огромный объем информации, собираемой различными датчиками, превышает возможности анализа человеком или группой людей. Учитывая постоянно растущий объем данных, доступных сегодня в мире, и повышение производительности обработки данных ожидается, что популярность ИИ будет продолжать расти.

Улучшенный таргетинг и обзор обстановки – ещё одно явное преимущество. Одной из областей, где перегрузка данными ощущается наиболее остро, является обработка изображений. Количество камер, осуществляющих наблюдение, будет продолжать расти. Существует потребность для автоматизации процесса анализа входящего потока видеоизображений. Возможности автоматизированного распознавания изображений и обнаружения объектов превзошли человеческие возможности, по крайней мере, в некоторых случаях. По мере дальнейшего прогресса эти системы будут все больше и больше способны идентифицировать объекты, которые люди пропустили бы. Это уже было выявлено в области медицины. Кроме того, прогресс в области распознавания лиц может быть применен для быстрой идентификации личностей, а анализ выражений лиц может предупредить о рискованных ситуациях.

Поддержка принятия решений в некоторых случаях сможет предоставить лучшие варианты выбора, чем люди могли бы предложить. Известным примером является технология маршрутизации, которая может получать полные карты и информацию о дорожном движении в реальном времени или проецируемую так, как это было бы недоступно людям.

Существует потенциал для применения ИИ в задачах стратегического планирования. Даже если эти технологии не подходят для использования при выработке боевых предложений или решений, эксперты ожидают, что они могут быть использованы для обеспечения более широкого спектра возможных

действий противника на учениях или оказать помощь в обнаружении грубых ошибок.

Смягчение кадровых проблем возможно, когда существует разрыв между спросом и персоналом, доступным для таких задач, как анализ изображений и перевод с иностранного языка. Это типы задач, которые возникают из-за быстрого роста объема данных, доступных для обработки. ИИ также играет ключевую роль в оказании роботизированной помощи на поле боя, что позволит поддерживать или расширять боевые возможности без увеличения численности личного состава.

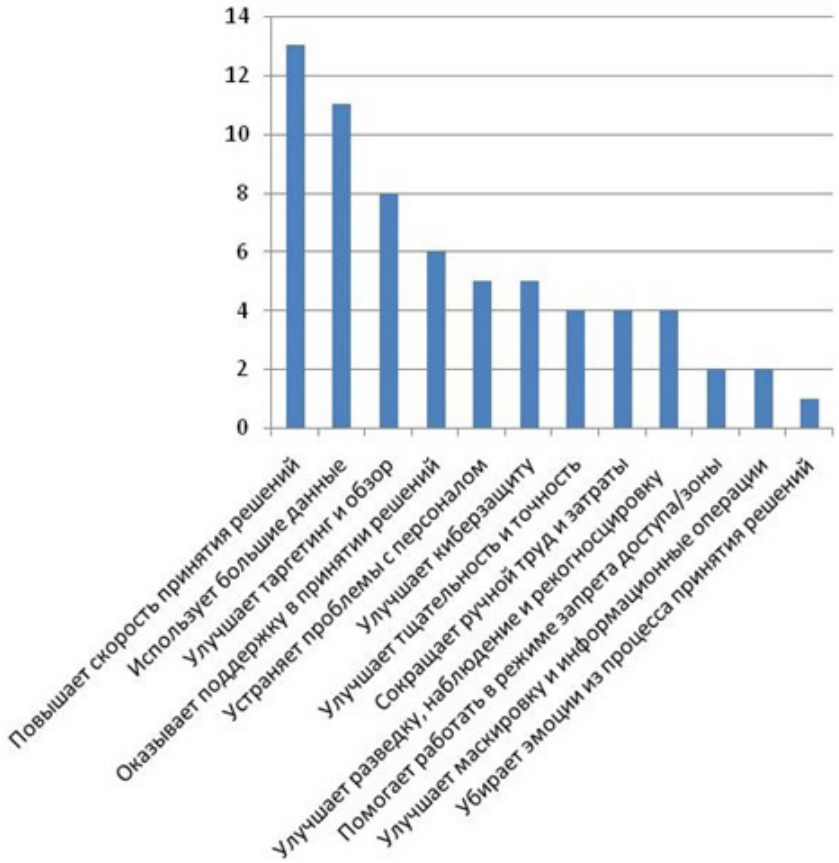

Рисунок 3 – Преимущества применения ИИ в области вооружений

Улучшения в киберзащите нужны, чтобы создать ИИ, который может наблюдать за программным обеспечением в системе и помечать действия, которые идентифицируются как подозрительные. Растет интерес к новым системам, которые могут находить и исправлять уязвимости в дружественных системах или находить и атаковать уязвимости во вражеских системах.

Повышение точности предоставляет преимущество по сравнению с людьми, которые, зачастую, привыкли использовать округлённые числа и грубые количественные оценки. Кроме того, операторы сложной техники подвержены повышенной утомляемости.

Снижение трудозатрат и затрат материальных средств позволяют сократить численность персонала без ущерба для услуг в диапазоне от логистики, транспорта до отопления и охлаждения помещений.

Улучшения в разведке, наблюдении и рекогносцировке (*ISR*) за счёт автономного сбора разведывательных данных с помощью дронов, с датчиков в наземной области, в космосе и даже в киберпространстве обещает еще больше увеличить объем генерируемых данных. И этот объем данных необходимо будет анализировать частично или полностью с помощью машин с использованием ИИ. Часть этого анализа необходимо будет выполнить на платформах *ISR*, развернутых в полевых условиях, из-за ограничений пропускной способности, которые делают невозможным передачу таких больших объемов данных. Большая часть анализа будет проводиться в центрах обработки разведывательных данных. Везде, где это делается, ИИ позволит значительно улучшить качество разведданных, полученных из массивов собранных данных *ISR*.

Способность работать в средах с запретом доступа в охраняемой зоне делает возможным уменьшение количества людей-операторов, подвергающихся риску в этих средах, но также могут быть меньше, быстрее и маневреннее, чем обитаемые оружейные платформы, и, следовательно, потенциально более боеспособны.

Улучшения в информационных операциях и создании дезинформации высокой реалистичности

является "преимуществом", подвергающимся сомнению респондентами в общественных опросах в силу возможных нарушений этических норм.

Устранение фактора эмоций и усталости из процесса принятия решений относится тоже к предполагаемым преимуществам.

Не меньший интерес представляют риски, выявленные в процессе опросов и интервью со специалистам. Чтобы оценить, будет ли использование ИИ в области вооружений разумным политическим выбором, необходимо сопоставить ожидаемые преимущества этих возможностей с рисками, которые они представляют. На рисунке 4 приведены опасения экспертов, отсортированные по частоте упоминания.

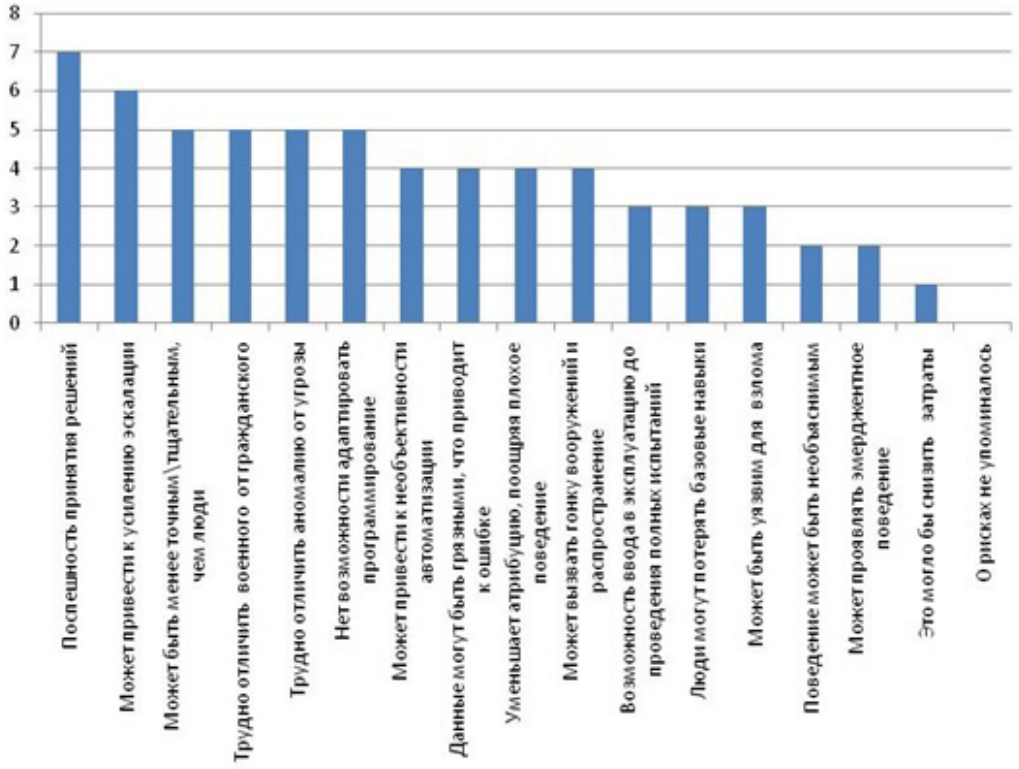

Рисунок 4 – Риски использования ИИ в области вооружений

Далее даны пояснения к наиболее распространённым упоминаниям рисков.

Хотя опрошенные эксперты назвали повышенную скорость, точность потенциальными преимуществами военного ИИ, они также выразили обеспокоенность тем, что эти возможности могут принимать решения слишком быстро или что системы не смогут адаптироваться к неизбежным сложностям военной обстановки. В результате они могут быть не в состоянии точно различать военных и мирных жителей или угрозы и системные аномалии, и в конечном итоге они могут быть менее точными и точными, чем люди-операторы. Эти проблемы могут усугубиться, если системы будут введены в эксплуатацию до того, как они будут должным образом протестированы, или в случае, если злоумышленникам удастся их взломать.

Искусственный интеллект может вызвать гонку вооружений или эскалацию конфликта. Автономное оружие может быть недостаточно чувствительным к политическим соображениям или порогам эскалации. Возможность того, что ИИ может снизить издержки войны с точки зрения человеческих жертв, могут побудить командиров идти на больший риск и действовать более агрессивно, что еще больше усиливает динамику эскалации.

Военные операторы и командиры могут слишком доверять своим системам искусственного интеллекта. Они могут проявлять «предвзятость к автоматизации», полагаясь на результаты этих систем ИИ, даже если они кажутся бессмысленными. Эта тенденция усиливается в системах, в которых алгоритмическая обработка настолько сложна, что их результаты необъяснимы, то есть операторы не могут легко определить, почему их системы дают определенные ответы или ведут себя определенным образом.

Упрощение работы военного оператора способствует снижению уровня его компетенций, поощряет его дисквалификацию.

Обсуждение. Военные приложения ИИ и другие автономные системы ответственного применения развиваются ускоренными темпами. Некоторые из упомянутых выше преимуществ уже реализованы в развернутых в настоящее время системах. Другие преимущества были продемонстрированы в контролируемых приложениях или лабораторных условиях. Третьи ожидаются на основе прогнозов о том, что обеспечит будущий прогресс в области ИИ, или экстраполяции того, какие военные приложения

можно создать на базе технологий, которые были продемонстрированы или разрабатываются в коммерческом секторе [14-15]. Тем не менее, несмотря на все потенциальные преимущества военного ИИ, существуют значительные риски из области этики, а также представляющие потенциальный ущерб, причиняемый в процессе оперативного и стратегического управления. Поскольку системы искусственного интеллекта могут совершать опасные ошибки, было бы самонадеянным торопиться с неизбирательной разработкой, развертыванием и использованием этих возможностей без более тщательного изучения рисков. Чтобы оценить, будет ли использование ИИ в вооружениях разумным политическим выбором, необходимо сопоставить ожидаемые преимущества этих возможностей с рисками, которые они представляют.

Выводы. Необходимость более тщательного изучения рисков в областях ответственного использования ИИ и особенно военного использования, отмеченная в открытом письме Илона Маска и других специалистов [13], набрала 31810 подписей. В РФ опубликован Кодекс этики в сфере искусственного интеллекта [16], который охватывает этическую сторону разработки, внедрения и применения ИИ на этапах жизненного цикла и ещё не отражён в законодательстве. Направление работы, заложенное в основе данных документов, это совместная разработка и внедрение набора общих протоколов безопасности для усовершенствованного проектирования и разработки искусственного интеллекта, которые тщательно проверяются и контролируются независимыми внешними экспертами. Эти протоколы должны обеспечивать то, что системы, выполняющие их, с приемлемо высокой вероятностью безопасны. Разработка протоколов не означает паузу в совершенствовании ИИ в целом, а представляют собой шаг назад от рискованной гонки ко все более непредсказуемым моделям по типу "черного ящика".

Параллельно разработчики ИИ должны работать с уполномоченными представителями по этике и праву, чтобы ускорить разработку надежных систем управления ИИ. Результатом их сотрудничества должны стать: новые и эффективные регулирующие органы, занимающиеся вопросами ИИ; наблюдение и контроль за высокопроизводительными системами искусственного интеллекта и большими пулами вычислительных машин; системы идентификации и "водяных знаков", помогающие отличить настоящие данные от синтезированных, системы обнаружения утечки моделей; защищённая экосистема аудита и сертификации; законодательно определённая ответственность за вред, причиненный ИИ; достаточное госбюджетное финансирование технических исследований безопасности ИИ.

СПИСОК ЛИТЕРАТУРЫ:

1. Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su And Moa Peldán Carlsson (2020). Artificial Intelligence, Strategic Stability and Nuclear Risk. https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

2. Integrating Cybersecurity and Critical Infrastructure. National, Regional and International Approaches. Edited by Lora Saalman (2018). https://www.sipri.org/sites/default/files/2018-04/integrating_cybersecurity_0.pdf.

3. The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Volume I. Euro-Atlantic Perspectives. Edited by Vincent Boulanin (2020). https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk

4. Vincent Boulanin and Maaike Verbruggen (2020). Mapping the Development of Autonomy in Weapon Systems. https://www.sipri.org/sites/default/files/2018-04/integrating_cybersecurity_0.pdf.

5. Vincent Boulanin, Laura Bruun and Netta Goussac (2021). Autonomous Weapon Systems And International Humanitarian Law. Identifying Limits and the Required Type and Degree of Human–Machine Interaction. https://www.sipri.org/sites/default/files/2021-06/2106_aws_and_ihl_0.pdf.

6. Lora Saalman, Fei Su and Larisa Saveleva Dovgal (2022). Cyber Posture Trends in China, Russia, the United States and the European Union. https://www.sipri.org/sites/default/files/2022-12/2212_cyber_postures_0.pdf.

7. Vincent Boulanin (2017). Mapping the development of autonomy in weapon systems. A primer on autonomy. https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf.

8. Vincent Boulanin, Netta Goussac, Laura Bruun And Luke Richards (2020). Responsible Military Use of Artificial Intelligence. Can the European Union Lead the Way in Developing Best Practice? https://www.sipri.org/publications/2020/other-publications/responsible-military-use-artificial-intelligence-can-european-union-lead-way-developing-best.

9. Vincent Boulanin, Kolja Brockmann And Luke Richards (2020). Responsible Artificial Intelligence Research And Innovation For International Peace And Security. https://www.sipri.org/sites/default/files/2020-11/sipri_report_responsible_artificial_intelligence_research_and_innovation_for_international_peace_and_security_2011.pdf.

10. Mark Bromley and Giovanna Maletta (2018). The Challenge of Software and Technology Transfers to Non-Proliferation Efforts. Implementing and Complying with Export Controls. https://www.sipri.org/publications/2018/other-publications/challenge-software-and-technology-transfers-non-proliferation-efforts-implementing-and-complying.

11. Johan Turell, Fei Su And Vincent Boulanin (2020). Cyber-incident Management Identifying and Dealing with the Risk of Escalation. IPRI Policy Paper N. 55. https://www.sipri.org/publications/2020/sipri-policy-papers/cyber-incident-management-identifying-and-dealing-risk-escalation.

12. Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, Derek Grossman. Military Applications of Artificial Intelligence. RAND Corporation, Santa Monica, Calif., 2020, Pages 202.

13. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

14. Михеев М.Ю. Моделирование и трансфер агротехнологий на основе системы поддержки принятия решений/ Михеев М.Ю., Прокофьев О.В., Савочкин А.Е., Сёмочкина И.Ю.// Труды международного симпозиума "Надежность и качество". - Пенза: Издательство Пензенского государственного университета, 2021. – Т. 1. – С. 180-183.

15. Михеев М.Ю. Методологии построения систем поддержки принятия решений в многоаспектной области применения/ Михеев М.Ю., Прокофьев О.В., Сёмочкина И.Ю. // Труды международного симпозиума "Надежность и качество". – Пенза: Издательство Пензенского государственного университета, 2022. – Т. 1. – С. 18-22.

16. Кодекс этики в сфере ИИ. URL: https://ethics.a-ai.ru/ (дата обращения: 30.06.2023).

Статья поступила в редакцию 29.06.2023
Статья принята к публикации 15.09.2023