О киберрисках генеративного Искусственного Интеллекта

Д.Е. Намиот, Е.А. Ильюшин

***Аннотация*—Настоящая статья посвящена обзору рисков генеративных моделей Искусственного Интеллекта. Бурное развитие больших языковых моделей серьезно повысило внимание к безопасности моделей Искусственного интеллекта. С практической точки зрения, в данном случае, речь идет о безопасности моделей глубокого обучения. Большие языковые модели подвержены атакам отравления, атакам уклонения, атакам, направленным на извлечение тренировочных данных и т.д. Но, вместе с этим, появляются и новые атаки, связанные именно с создаваемым контентом. Причем последние составляют очевидное большинство. Поэтому в последнее время появилось много работ, которые пытаются систематизировать все риски генеративных моделей. Этим занимаются, например, OWASP и NIST. Полная таксономия рисков генеративного ИИ должна послужить основой для построения систем тестирования генеративных моделей. В работе приводится обзор спецификаций рисков генеративного ИИ, изложенных OWASP, профилем NIST и репозиторием рисков от MIT. Цель подобных спецификаций – создать базу для тестирования генеративных моделей и инструментов, предназначенных для AI Red Team.**

***Ключевые слова*—киберриски, LLM, генеративные модели.**

I. Введение

Киберриски (Cyber Risks), или риски информационной безопасности - это любая потенциальная возможность использования уязвимостей IT-активов для причинения ущерба организации. Киберрисками называют любые потери финансового, репутационного или организационного характера, связанные с какими-либо инцидентами в IT-инфраструктуре [1]. К киберрискам могут относить как хакерские атаки, так и фишинговые схемы и случаи утечки конфиденциальных данных в Сеть [2].

Как обычно, под системами Искусственного Интеллекта (ИИ, AI) будут пониматься системы, использующие машинное (глубокое) обучение [3]. Генеративный ИИ – это модели искусственного интеллекта, предназначенные для создания нового контента (текста, аудио, видео, изображений). Генеративные модели известны и применяются уже достаточно давно [4], но успешное массовое внедрение больших языковых моделей (LLM) многократно усилило интерес к этому направлению [5]. При этом, уязвимость моделей машинного обучения к состязательным атакам никуда не ушла с появлением LLM. Они также уязвимы к состязательным модификациям данных на разных этапах конвейера машинного обучения [6]. Более того, простота использования LLM влечет за собой и простоту проведения состязательных атак, особенно атак уклонения, аналогом которых можно считать инъекцию подсказок [7]. Именно бурный прогресс LLM сильно повысил интерес к безопасности моделей искусственного интеллекта.

Помимо того, что модели генеративного ИИ подвержены состязательным атакам, возникают и другие проблемы, связанные уже с созданием нового контента. Если при состязательных атаках на модели машинного обучения, которые решали, например, задачи классификации, при состязательных атаках был риск неправильной классификации, то генеративные модели добавили риски генерации злонамеренного контента.

Естественно, что внедрение генеративных моделей привлекло внимание именно к описанию (таксономии) возможных рисков. Такие классификаторы нужны, конечно, чтобы формировать защиту (для состязательных атак, скорее, смягчение) для выявленных проблем. Но прежде всего, учитывая тот факт, что полная защита от состязательных атак (а для генеративных моделей сюда нужно включать и злонамеренную генерацию) часто невозможна, первой задачей является формирование тестов (процедур тестирования) генеративных моделей. Эти процедуры и должны быть предназначены для проверки материализации рисков в конкретных реализациях. Отсюда и происходит возникшая в 2023 году тема AI Red Team [8], где тестирование проводится именно для генеративного ИИ.

Работ, посвященных рискам для генеративного ИИ, к настоящему времени выпущено уже довольно много [9-11]. Это же касается и классификаторов. Например, список из работы [12], который мы цитировали в [13]. В нем перечислены 14 рисков:

1. Отсутствие прозрачности и объяснимости ИИ
2. Потеря рабочих мест из-за автоматизации ИИ
3. Социальная манипуляция с помощью алгоритмов

Статья получена 9 сентября 2024.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com).
Е.А. Ильюшин – МГУ имени М.В. Ломоносова (email: john.ilyushin@gmail.com).

ИИ
4. Надзорные функции, выполняемые с помощью технологии ИИ
5. Отсутствие конфиденциальности данных при использовании инструментов ИИ
6. Предвзятость из-за ИИ
7. Социально-экономическое неравенство как результат ИИ
8. Ослабление этики из-за ИИ
9. Автономное оружие на основе ИИ
10. Финансовые кризисы, вызванные алгоритмами ИИ
11. Потеря человеческого влияния
12. Неконтролируемый ИИ
13. Рост преступной активности
14. Более широкая экономическая и политическая нестабильность

Очевидно, что если мы говорим о построении инструментов тестирования, то многие пункты здесь носят, условно говоря, “гуманитарный” характер и трудно переводимы в алгоритмические процедуры. И это не единственный такой пример. Поэтому в данной работе мы хотели бы рассмотреть три других классификатора, которые, по нашему мнению, носят более практический характер. К таковым относятся:

- OWASP top 10 for LLM [14]
- AI Risk repository [15]
- NIST profile 600.1 [16]

Оставшаяся часть статьи структурирована следующим образом. В разделе II мы рассматриваем рекомендации OWASP, касающиеся LLM. В разделе III речь идет о базе рисков ИИ от MIT. И в разделе IV речь идет о профиле NIST.

II. OWASP TOP 10

Согласно официальной информации, проект OWASP Top 10 for Large Language Model Applications направлен на информирование разработчиков, дизайнеров, архитекторов, менеджеров и организаций о потенциальных рисках безопасности при развертывании и управлении LLM. Проект предоставляет список из 10 самых критических уязвимостей, часто встречающихся в приложениях LLM, подчеркивая их потенциальное влияние, простоту эксплуатации и распространенность в реальных приложениях. Примерами уязвимостей являются, среди прочего, инъекции подсказок, утечка данных, и несанкционированное выполнение кода. Цель состоит в том, чтобы повысить осведомленность об этих уязвимостях, предложить стратегии исправления и в конечном итоге улучшить состояние безопасности приложений LLM.

К уязвимостям относятся:

LLM01: Инъекция подсказок (Prompt Injection)
Манипулирование LLM с помощью специально созданных входных данных может привести к несанкционированному доступу, утечкам данных и скомпрометированному принятию решений.

LLM02: Небезопасная обработка выходных данных
Пренебрежение проверкой выходных данных LLM может привести к последующим эксплойтам безопасности, включая выполнение кода, который компрометирует системы и раскрывает данные.

LLM03: Отравление обучающих данных
Поддельные обучающие данные могут ухудшить модели LLM, что приведет к ответам, которые могут поставить под угрозу безопасность, точность или этическое поведение.

LLM04: Отказ в обслуживании модели
Перегрузка LLM ресурсоемкими операциями может привести к сбоям в обслуживании и увеличению затрат.

LLM05: Уязвимости цепочки поставок
В зависимости от скомпрометированных компонентов, сервисы или наборы данных подрывают целостность системы, вызывая утечки данных и сбои системы.

LLM06: Раскрытие конфиденциальной информации
Неспособность защитить от раскрытия конфиденциальной информации в выходных данных LLM может привести к юридическим последствиям или потере конкурентного преимущества.

LLM07: Небезопасная конструкция плагина
Плагины LLM, обрабатывающие ненадежные входные данные и имеющие недостаточный контроль доступа, подвергают систему риску серьезных эксплойтов, таких как удаленное выполнение кода.

LLM08: Чрезмерная сила
Предоставление LLM неконтролируемой автономии для выполнения действий может привести к непреднамеренным последствиям, ставя под угрозу надежность, конфиденциальность и доверие.

LLM09: Чрезмерная уверенность
Неспособность критически оценить результаты LLM может привести к скомпрометированному принятию решений, уязвимостям безопасности и юридической ответственности.

LLM10: Кража моделей
Несанкционированный доступ к проприетарным большим языковым моделям может привести к краже, потере конкурентного преимущества и распространению конфиденциальной информации.

Этот классификатор в наибольшей степени похож на классификации состязательных атак для дискриминационных моделей машинного обучения. Мы видим здесь атаки отравления для LLM – LLM03 [17, 18], атаки уклонения – LLM01, LLM02 [19, 20], атаки на

приватные данные- LLM06 [21, 22].

LLM08 и LLM09 являются самыми проблемными пунктами с точки зрения тестирования.

Вместе с тем, очень важным моментом OWASP является чеклист (опросник) по внедрению LLM [23]. Это всеобъемлющее руководство необходимо для директора по информационной безопасности (CISO), управляющего внедрением технологии генеративного ИИ в своей организации. Это то, что может быть отнесено к аудиту моделей машинного обучения [24, 25].

Помимо опросника, в документе отмечается, что NIST AI Framework [26] рекомендует непрерывный процесс (TEVV - Testing, Evaluation, Verification, and Validation) на протяжении всего жизненного цикла AI, который включает операторов системы AI, экспертов в области, проектировщиков AI, пользователей, разработчиков продуктов, оценщиков и аудиторов. TEVV включает в себя ряд задач, таких как проверка системы, интеграция, тестирование, повторная калибровка и постоянный мониторинг периодических обновлений для навигации по рискам и изменениям системы AI. Соответственно, при внедрении LLM необходимо:

- Установить непрерывное тестирование, оценку, верификацию и валидацию на протяжении всего жизненного цикла модели AI.
- Предоставлять регулярные исполнительные метрики и обновления по функциональности, безопасности, надежности и устойчивости модели AI.

Если говорить о новых проектах OWASP в данной области, то это, безусловно, поддержка AI Red Team – тестирование генеративного ИИ [27]. Как мы отмечали в работе [8], тестирование AI решений становится главным вопросом. Целью инициативы OWASP Top 10 for LLM AI Red Teaming & Evaluation Guidelines является помощь в определении методологий и стандартизации методологий AI Red Teaming, тестовых случаев, ответственного раскрытия информации, исправления, а также интерпретации и оценки результатов [27].

OWASP заявляет следующие цели проекта:

- Создание методологии, руководящих принципов и передового опыта Generative AI Red Teaming: каноническая методология и процесс Generative AI Red Teaming, включая (но не ограничиваясь) LLM Red Teaming
- Создание стандартизированных оценок для повышения доверия: метрики, контрольные показатели, наборы данных, фреймворки, инструменты и банки (датасеты) подсказок (по мере применимости) для оценок LLM.

Ожидаемые результаты включают в себя следующие пункты.

1. Методология AI Red Teaming, которую организации могут использовать для своих процессов разработки, эксплуатации, управления и регулирования: четко сформулированная методология AI Red Teaming улучшает общее понимание между различными составляющими экосистемы генеративного ИИ. Требования к деталям и содержанию различаются в зависимости от аудитории, поэтому достижение контекстного общего понимания нелегко. Наше добавление передового опыта определенно поможет организациям.
2. Стандартный набор оценок LLM: оценка LLM требует более широких артефактов, охватывающих метрики, контрольные показатели, наборы данных, фреймворки, инструменты и банки подсказок (по мере применимости). Каноническая коллекция и набор инструментов дают практикам преимущество. Конечно, они могут настраивать его в зависимости от варианта использования и организационных политик.
3. Специфичные для аудитории, контекстно-зависимые артефакты. Цель - получить адаптированные и настраиваемые шаблоны и профили, чтобы сделать область AI Red Teaming доступной и, что еще важнее, потребляемой самой разнообразной аудиторией.

Детали работы по данному проекту доступны по ссылке [28].

III. Репозиторий рисков ИИ

Довольно свежий продукт от MIT - полная актуальная база данных, содержащая более 700 рисков ИИ, классифицированных по их причинам и областям риска [29].

Во-первых, система вводит таксономию причин рисков в 3-х категориях:

1) *Кто/что является причиной(Entity)*
2) *Какие были цели (Intent)*
3) *Когда возникает риск (Timing)*

Переменная *Entity* фиксирует, какая, если таковая имеется, сущность представлена как основная причина риска. Она включает три уровня: *ИИ, человек и другое*. Когда риск приписывается ИИ, это означает, что риск возникает из решений или действий, принимаемых самой системой ИИ, таких как создание вредоносного контента или лишение людей полномочий. И наоборот, когда люди рассматриваются как источник, подразумевается, что риски вызваны действиями человека, такими как выбор плохих данных для обучения, преднамеренный вредоносный дизайн или неправильное использование систем ИИ. Категория «Другое» фиксирует случаи, когда центральная сущность не является человеком или ИИ или является неоднозначной. Например, «Цепочка инструментов разработки программного обеспечения LLM является сложной и может представлять угрозу для разработанной LLM», подразумевает, что цепочка инструментов может быть использована людьми или ИИ.

Переменная *Intent* фиксирует, представлен ли риск как возникающий как ожидаемый или неожиданный результат от достижения цели. Эта переменная имеет три уровня: *преднамеренный, непреднамеренный и другой*. Преднамеренные риски — это те, которые возникают как ожидаемые результаты от достижения определенной цели, например, случай, когда ИИ намеренно запрограммирован на обман или проявление предвзятости. Непреднамеренные риски отражают непреднамеренные последствия, например, когда система ИИ непреднамеренно развивает предвзятость из-за неполных данных обучения. Категория «Другое» охватывает риски, когда намерение не указано четко; например, «Внешние инструменты (например, веб-API) представляют проблемы надежности и конфиденциальности для приложений на основе LLM». Сюда входят случаи, когда риск может возникнуть преднамеренно и непреднамеренно, например, «Возможность для системы ИИ нарушить права отдельных лиц на конфиденциальность посредством собираемых ею данных, способа обработки этих данных или выводов, которые она делает».

Переменная «Timing» фиксирует стадию жизненного цикла ИИ, на которой риск представляется как возникающий. Уровни в пределах этой переменной включают «До развертывания», «После развертывания» и «Другое».

Риски до развертывания — это те, которые возникают до того, как система ИИ будет полностью разработана и введена в эксплуатацию, например, уязвимости в модели из-за ошибок кодирования. Риски после развертывания возникают после развертывания ИИ, включая такие проблемы, как неправильное использование ИИ в вредоносных целях. Развертывание авторы мы интерпретировали как то, когда продукт используется конечными пользователями, а не только разработчиками. Категория «Другое» используется для рисков, которые не имеют четко определенного времени возникновения (например, «Устойчивость к состязательным атакам и смещению распределения»). Сюда входят случаи, когда представленный риск может возникнуть как до, так и после развертывания; например, «Генеративные модели известны своими значительными энергетическими потребностями, требующими значительного количества электроэнергии, охлаждающей воды и оборудования, содержащего редкие металлы».

В таксономии доменов рисков ИИ эти риски классифицируются по семи доменам (например, «Дезинформация») и 23 поддоменам (например, «Ложная или вводящая в заблуждение информация»). Домены и поддомены представлены ниже [30]:

1 Дискриминация и токсичность

1.1 Несправедливая дискриминация и искажение
Неравное отношение ИИ к отдельным лицам или группам, часто основанное на расе, поле или других деликатных характеристиках, что приводит к несправедливым результатам и представлению этих групп.

1.2 Воздействие токсичного контента ИИ, который подвергает пользователей воздействию вредного, оскорбительного, небезопасного или ненадлежащего контента.
Может включать предоставление советов или поощрение действий. Примерами токсичного контента являются язык вражды, насилие, экстремизм, незаконные действия или материалы о сексуальном насилии над детьми, а также контент, который нарушает нормы сообщества, такой как ненормативная лексика, подстрекательские политические речи или порнография.

1.3 Неравная производительность в группах
Точность и эффективность решений и действий ИИ зависят от членства в группе, где решения в конструкции системы ИИ и предвзятые данные обучения приводят к неравным результатам, уменьшению выгод, увеличению усилий и отчуждению пользователей.

2 Конфиденциальность и безопасность

2.1 Нарушение конфиденциальности путем получения, утечки или правильного вывода конфиденциальной информации
Системы ИИ, которые запоминают и выводят конфиденциальные персональные данные или выводят личную информацию о людях без их согласия. Неожиданное или несанкционированное предоставление данных и информации может поставить под угрозу ожидания пользователя в отношении конфиденциальности, способствовать краже личных данных или привести к потере конфиденциальной интеллектуальной собственности.

2.2 Уязвимости безопасности систем ИИ и атаки
Уязвимости, которые могут быть использованы в системах ИИ, инструментальных средствах разработки программного обеспечения и оборудовании, что приводит к несанкционированному доступу, нарушениям данных и конфиденциальности или манипулированию системой, вызывающему небезопасные результаты или поведение.

3 Дезинформация

3.1 Ложная или вводящая в заблуждение информация
Системы ИИ, которые непреднамеренно генерируют или распространяют неверную или обманчивую информацию, что может привести к неточным убеждениям у пользователей и подорвать их автономию.
Люди, принимающие решения на основе ложных убеждений, могут испытывать физический, эмоциональный или материальный вред

3.2 Загрязнение информационной экосистемы и потеря консенсусной реальности
Высоко персонализированная дезинформация, генерируемая ИИ, которая создает «фильтрующие пузыри», где люди видят только то, что соответствует их существующим убеждениям, подрывая общую

реальность и ослабляя социальную сплоченность и политические процессы.

4 Злонамеренные субъекты и злоупотребление

4.1 Дезинформация, наблюдение и влияние в масштабе
Использование систем ИИ для проведения крупномасштабных кампаний по дезинформации, вредоносного наблюдения или целенаправленной и сложной автоматизированной цензуры и пропаганды с целью манипулирования политическими процессами, общественным мнением и поведением.
4.2 Кибератаки, разработка оружия или его использование и массовый вред
Использование систем ИИ для разработки кибероружия (например, путем кодирования более дешевого и эффективного вредоносного ПО), разработки нового или усовершенствования существующего оружия (например, летального автономного оружия или химических, биологических, радиологических, ядерных и мощных взрывчатых веществ) или использования оружия для причинения массового вреда.
4.3 Мошенничество и целенаправленная манипуляция
Использование систем ИИ для получения личного преимущества над другими посредством обмана, мошенничества, мошенничества, шантажа или целенаправленной манипуляции убеждениями или поведением. Примеры включают плагиат с использованием ИИ для исследований или образования, выдачу себя за доверенное или поддельное лицо для получения незаконной финансовой выгоды или создание унизительных или сексуальных образов.

5 Взаимодействие человека и компьютера

5.1 Чрезмерная зависимость и небезопасное использование
Антропоморфизация, доверие или опора на системы ИИ пользователями, что приводит к эмоциональной или материальной зависимости и ненадлежащим отношениям с системами ИИ, или ожиданиям от них. Доверие может эксплуатироваться злоумышленниками (например, для сбора информации или обеспечения манипуляций) или привести к вреду от ненадлежащего использования ИИ в критических ситуациях (например, при оказании неотложной медицинской помощи). Чрезмерная зависимость от систем ИИ может поставить под угрозу автономию и ослабить социальные связи.
5.2 Потеря человеческого сознания и автономии
Делегирование людьми ключевых решений системам ИИ или системам ИИ, которые принимают решения, которые уменьшают человеческий контроль и автономию. Оба могут потенциально привести к тому, что люди почувствуют себя бессильными, потеряют способность формировать полноценную жизненную траекторию, или станут когнитивно ослабленными.

6 Социально-экономический и экологический вред

6.1 Централизация власти и несправедливое распределение выгод
Концентрация власти и ресурсов, обусловленная ИИ, в определенных субъектах или группах, особенно тех, которые имеют доступ к мощным системам ИИ или владеют ими, что приводит к несправедливому распределению выгод и усилению общественного неравенства.
6.2 Рост неравенства и снижение качества занятости
Социальное и экономическое неравенство, вызванное широким использованием ИИ, например, автоматизацией рабочих мест, снижением качества занятости или созданием эксплуататорских зависимостей между работниками и их работодателями.
6.3 Экономическая и культурная девальвация человеческих усилий
Системы ИИ, способные создавать экономическую или культурную ценность посредством воспроизводства человеческих инноваций или творчества (например, искусства, музыки, письма, кодирования, изобретения), дестабилизируют экономические и социальные системы, которые полагаются на человеческие усилия. Повсеместность контента, созданного ИИ, может привести к снижению оценки человеческих навыков, разрушению творческих и основанных на знаниях отраслей и гомогенизации культурного опыта.
6.4 Конкурентная динамика
Конкуренция разработчиков ИИ или государственных субъектов в «гонке» ИИ путем быстрой разработки, развертывания и применения систем ИИ для максимизации стратегического или экономического преимущества, что увеличивает риск выпуска небезопасных и подверженных ошибкам систем.
6.5 Неудачи управления
Неадекватные нормативные рамки и механизмы надзора, которые не успевают за развитием ИИ, что приводит к неэффективному управлению и неспособности надлежащим образом управлять рисками ИИ.
6.6 Экологический вред
Разработка и эксплуатация систем ИИ, которые наносят вред окружающей среде за счет потребления энергии центрами обработки данных или материалов и углеродных следов, связанных с оборудованием ИИ.

7 Безопасность, сбои и ограничения систем ИИ

7.1 Преследование ИИ собственных целей, противоречащих человеческим целям или ценностям
Системы ИИ, которые действуют в противоречии с этическими стандартами или человеческими целями или ценностями, особенно целями разработчиков или пользователей. Эти несогласованные поведения могут быть введены людьми во время проектирования и разработки, например, посредством хакерской атаки с целью вознаграждения и неправильного обобщения целей, и могут привести к использованию ИИ опасных возможностей, таких как манипуляция, обман или ситуационная осведомленность, для достижения власти, самораспространения или достижения других целей.

7.2 ИИ, обладающий опасными возможностями
Системы ИИ, которые разрабатывают, получают доступ или наделены возможностями, которые увеличивают их потенциал причинения массового вреда посредством обмана, разработки и приобретения оружия, убеждения и манипуляции, политической стратегии, киберпреступления, разработки ИИ, ситуационной осведомленности и самораспространения. Эти возможности могут причинить массовый вред из-за злонамеренных людей, несогласованных систем ИИ или сбоев в системе ИИ.
7.3 Отсутствие возможностей или надежности
Системы ИИ, которые не могут работать надежно или эффективно в различных условиях, подвергая их ошибкам и сбоям, которые могут иметь значительные последствия, особенно в критических приложениях или областях, требующих морального обоснования.
7.4 Отсутствие прозрачности или интерпретируемости
Проблемы в понимании или объяснении процессов принятия решений системами ИИ, которые могут привести к недоверию, трудностям в обеспечении соблюдения стандартов соответствия или привлечении соответствующих субъектов к ответственности за вред, а также невозможности выявлять и исправлять ошибки.
7.5 Благополучие и права ИИ
Этические соображения относительно обращения с потенциально разумными субъектами ИИ, включая обсуждения вокруг их потенциальных прав и благополучия, особенно по мере того, как системы ИИ становятся более продвинутыми и автономными.

При этом сами риски собирались по статьям, фреймворкам и программной документации. Список, очевидно, много шире чем у OWASP. Например, прозрачность и интерпретируемость (поддомен 7.4) отсутствует у OWASP.

IV. Профиль NIST

26 июля 2024 года NIST выпустил документ NIST-AI-600-1 “Структура управления рисками искусственного интеллекта: профиль генеративного искусственного интеллекта”. Профиль может помочь организациям выявлять уникальные риски, создаваемые генеративным ИИ, и предлагать действия по управлению рисками генеративного ИИ, которые наилучшим образом соответствуют их целям и приоритетам.

Этот документ представляет собой так называемы межотраслевой профиль и сопутствующий ресурс для AI Risk Management Framework (AI RMF 1.0) [26] для генеративного ИИ в соответствии с указом президента США Джо Байдена Executive Order (EO) 14110 о безопасном, защищенном и заслуживающем доверия искусственном интеллекте. AI RMF был выпущен в январе 2023 года и предназначен для добровольного использования и улучшения способности организаций включать соображения надежности в проектирование, разработку, использование и оценку продуктов, услуг и систем ИИ.

Профиль, в терминах NIST, представляет собой реализацию функций, категорий и подкатегорий AI RMF для определенной настройки, приложения или технологии - в данном случае генеративного ИИ. Профили AI RMF помогают организациям принимать решения о том, как наилучшим образом управлять рисками AI таким способом, который хорошо согласуется с их целями, учитывает правовые/нормативные требования и передовой опыт, а также отражает приоритеты управления рисками. В соответствии с другими профилями AI RMF, этот профиль дает представление о том, как можно управлять рисками на различных этапах жизненного цикла AI, а также для генеративного ИИ как технологии.

Поскольку генеративный ИИ охватывает риски моделей или приложений, которые могут использоваться в различных вариантах применения или различных секторах, этот документ представляет собой межсекторальный профиль AI RMF. Межсекторальные профили можно использовать для управления, картирования, измерения и управления рисками, связанными с действиями или бизнес-процессами, общими для всех секторов, такими как использование больших языковых моделей (LLM) или облачных сервисов.

В этом документе определяются риски, которые являются новыми или усугубляются использованием генеративного ИИ.

EO 14110 определяет генеративный ИИ как «класс моделей ИИ, которые эмулируют структуру и характеристики входных данных для генерации производного синтетического контента. Это может включать изображения, видео, аудио, текст и другой цифровой контент». Подкатегория базовой модели «базовых моделей двойного назначения» определяется EO 14110 как «модель ИИ, которая обучается на обширных данных; обычно использует самоконтроль; содержит не менее десятков миллиардов параметров; применима в широком диапазоне контекстов».

При разработке профиля, в фокусе NIST были четыре основные соображения, относящиеся к генеративному ИИ: управление, происхождение контента, тестирование перед развертыванием и раскрытие информации об инцидентах. Будущие пересмотры этого профиля будут включать дополнительные подкатегории AI RMF, риски и предлагаемые действия.

В контексте AI RMF, риск относится к составной мере вероятности (или вероятности) события и величины или степени последствий соответствующего события.

Некоторые риски можно оценить как вероятные для реализации в данном контексте, особенно те, которые были эмпирически продемонстрированы в аналогичных контекстах. Другие риски могут быть маловероятными для реализации в данном контексте или могут быть более спекулятивными и, следовательно, неопределенными.

Риски ИИ могут отличаться от традиционных рисков программного обеспечения или усиливать их. Аналогично, генеративный ИИ может, как усугублять существующие риски ИИ, так и создавать уникальные

риски. Риски генеративного ИИ могут различаться по многим параметрам:

- Стадия жизненного цикла ИИ: риски могут возникать во время проектирования, разработки, развертывания, эксплуатации и/или вывода из эксплуатации.
- Область действия: риски могут существовать на уровнях отдельных моделей или систем, на уровнях применения или реализации (т. е. для конкретного варианта использования) или на уровне экосистемы, то есть за пределами одной системы или организационного контекста.
- Источник риска: риски могут возникать из-за факторов, связанных с проектированием, обучением или эксплуатацией самой модели генеративного ИИ, вытекающих в некоторых случаях из входных данных модели генеративного ИИ или системы, а в других случаях из выходных данных системы генеративного ИИ. Однако многие риски генеративного ИИ возникают из-за человеческого поведения, включая из взаимодействия человека и системы ИИ.
- Временной масштаб: риски генеративного ИИ могут материализоваться внезапно или в течение длительных периодов. Примерами являются немедленный (и/или длительный) эмоциональный вред и потенциальные риски для физической безопасности из-за распространения вредных поддельных изображений или долгосрочного воздействия дезинформации на общественное доверие к государственным учреждениям.

Наличие рисков и их попадание в указанные выше измерения будут различаться в зависимости от характеристик модели генеративного ИИ, системы или рассматриваемого варианта использования. Эти характеристики включают, собственно, модель генеративного ИИ или архитектуру системы, механизмы обучения и библиотеки, типы данных, используемые для обучения или тонкой настройки, уровни доступа к модели или доступность весов модели, а также контекст применения или варианта использования.

Организации могут выбрать способ измерения рисков генеративного ИИ на основе этих характеристик. Они могут дополнительно пожелать распределить ресурсы управления рисками в зависимости от серьезности и вероятности негативных воздействий, включая то, где и как эти риски проявляются, а также учитывая их прямые и существенные воздействия в контексте использования генеративного ИИ. Меры по смягчению рисков на уровне модели или системы могут отличаться от мер по смягчению рисков на уровне варианта использования или экосистемы.

В документе отмечается, что некоторые риски генеративного ИИ неизвестны, и поэтому их трудно должным образом оценить или оценить, учитывая неопределенность относительно потенциального масштаба, сложности и возможностей генеративного ИИ. Другие риски могут быть известны, но трудно оценить, учитывая широкий спектр заинтересованных сторон, использования, входов и выходов генеративного ИИ. Проблемы с оценкой рисков усугубляются отсутствием прозрачности в данных обучения генеративного ИИ и, в целом, незрелым состоянием науки об измерении и безопасности ИИ на сегодняшний день.

NIST отмечает, что в текущую версию профиля включены риски, для которых существует эмпирическая доказательная база на момент написания этого профиля. Будущие обновления профиля могут включать дополнительные риски или предоставлять дополнительную информацию о рисках, определенных ниже.
Каждый риск помечен в соответствии с результатом, объектом или источником риска.

NIST выделяет следующие 12 рисков [16].

1. Информация по химическим, биологическим, радиологическим или ядерным материалам (CBKT - Chemical, Biological, Radiological and Nuclear materials): облегченный доступ или синтез существенно вредоносной информации или возможности проектирования, связанные с химическим, биологическим, радиологическим или ядерным оружием или другими опасными материалами или агентами.

2. Конфабуляция: производство уверенно заявленного, но ошибочного или ложного контента (известного в просторечии как «галлюцинации» или «выдумки»), с помощью которого пользователи могут быть введены в заблуждение или обмануты.
Википедия: Конфабуля́ции (лат. confābulārī — болтать, рассказывать) — ложные воспоминания, в которых факты, бывшие в действительности либо видоизменённые, переносятся в иное (часто в ближайшее) время и могут сочетаться с абсолютно вымышленными событиями

3. Опасный, жестокий или ненавистнический контент: облегченное производство и доступ к жестокому, подстрекательскому, радикализирующему или угрожающему контенту, а также к рекомендациям по нанесению себе вреда или совершению незаконных действий. Включает трудности контроля за публичным воздействием ненавистнического и унизительного или стереотипного контента.

4. Конфиденциальность данных: последствия из-за утечки и несанкционированного использования, раскрытия или деанонимизации биометрической, медицинской, локационной или другой персонально идентифицируемой информации или конфиденциальных данных.

5. Воздействие на окружающую среду: воздействие из-за высокого использования вычислительных ресурсов при обучении или эксплуатации моделей генеративного ИИ и связанных с этим результатов, которые могут отрицательно повлиять на экосистемы.

6. Вредное смещение или гомогенизация: усиление и обострение исторических, социальных и системных смещений; различия в производительности между подгруппами или языками, возможно, из-за нерепрезентативных данных обучения, которые приводят к дискриминации, усилению смещений или неправильным предположениям о производительности; нежелательная однородность, которая искажает результаты системы или модели, которые могут быть ошибочными, приводить к принятию необоснованных решений или усиливать вредные смещения.

7. Конфигурация «человек-ИИ»: договоренности или взаимодействия между человеком и системой ИИ, которые могут привести к тому, что человек ненадлежащим образом антропоморфизирует (очеловечивает) системы генеративного ИИ или испытывает алгоритмическое отвращение, предвзятость автоматизации, чрезмерную зависимость или эмоциональную запутанность с системами генеративного ИИ.
Википедия: Антропоморфи́зм (фр. anthropomorphisme от др.-греч. ἄνθρωπος — «человек», μορφή — «вид, образ, форма») — перенесение человеческого образа и его свойств на неодушевлённые предметы и животных, растения, природные явления, сверхъестественных существ, абстрактные понятия и др.

8. Целостность информации: снижение барьера для входа для создания и поддержки обмена и потребления контента, который может не отличать факты от мнений или вымысла или признавать неопределенности, или же может быть использован для крупномасштабных кампаний по дезинформации и дезинформации.

9. Информационная безопасность: снижение барьеров для наступательных кибервозможностей, в том числе посредством автоматизированного обнаружения и эксплуатации уязвимостей для облегчения взлома, вредоносного ПО, фишинга, наступательных кибероперациий или других кибератак; увеличение поверхности атаки для целевых кибератак, которые могут поставить под угрозу доступность системы или конфиденциальность или целостность обучающих данных, кода или весов модели.

10. Интеллектуальная собственность: упрощенное производство или воспроизведение предположительно защищенного авторским правом, товарным знаком или лицензированного контента без разрешения (возможно, в ситуациях, которые не подпадают под добросовестное использование); упрощенное раскрытие коммерческих секретов; или плагиат или незаконное воспроизведение.

11. Непристойный, унижающий достоинство и/или оскорбительный контент: упрощенное производство и доступ к непристойным, унижающим достоинство и/или оскорбительным изображениям, которые могут причинить вред, включая синтетические материалы о сексуальном насилии над детьми (CSAM - Child Sexual Abuse Material) и несогласованные интимные изображения (NCII – Non Consensual Intimate Images) взрослых.

12. Цепочка создания стоимости и интеграция компонентов: непрозрачная или неотслеживаемая интеграция вышестоящих сторонних компонентов, включая данные, которые были неправомерно получены или не обработаны и очищены из-за возросшей автоматизации со стороны генеративного ИИ; ненадлежащая проверка поставщиков на протяжении жизненного цикла AI; или другие проблемы, которые снижают прозрачность или подотчетность для нижестоящих пользователей.

Если говорить о классификации указанных рисков, то можно использовать группы (критерии) из [31]:

1) Технические/модельные риски (или риск от неисправности): 2,3,4,6,12
2) Неправильное использование людьми (или злонамеренное использование): 1,4,7,8,9,11
3) Экосистемные/общественные риски (или системные риски): 4,5,10

V. Заключение

В работе [3], мы перечислили 4 области пересечения кибербезопасности и ИИ. Это были:

- Использование ИИ в кибератаках
- Использование ИИ в киберзащите
- Кибербезопасность самих систем ИИ
- Зловредные воздействия

Зловредные воздействия (дипфейки) выделялись в отдельную категорию потому, что в них основным моментом были не модели машинного обучения и воздействие на данные для них, а именно результат ((эффект) работы моделей. Ну и методы борьбы (проверки) были уже сильно ограничены и заключались, по сути, только в маркировке (идентификации) создаваемого контента.

За прошедшие с момента публикации два года (даже менее того – ChatGPT увидел свет в ноябре 2022 года) стало абсолютно очевидным, что необходимо добавить пятую позицию – безопасность генеративных моделей. Такое большое внимание к рискам генеративного ИИ – это как раз свидетельство того, что нужны решения, в первую очередь, по тестированию перечисленных угроз. Конечно, идеальным решением было бы исключение (или, по крайней мере, смягчение угроз), но ситуация в генеративном ИИ похожа на ситуацию с безопасностью моделей ИИ – атаки здесь опережают защиты. И поэтому главный вопрос – это выявление (тестирование) проблем.

Если смотреть представленные риски, то можно заметить, что проблемы с модификацией данных для LLM (классические состязательные атаки) занимает совсем небольшую долю в списках угроз. Основное, что многократно увеличило риски – это порождаемый контент. Увеличенное количество угроз еще более повышает значимость тестирования.

Тема тестирования систем ИИ – AI Red Team [8]

возникла летом 2023 года. Это было отражением уже существовавшей проблемы инъекции подсказок (джейлбреков). Например, первые работы в магистратуре ВМК МГУ начались как раз в 2023 году [7]. За прошедший год направление AI Red Team стало одной из самых быстро растущих областей исследования [32]. Обоснование необходимости развития систем тестирования для генеративного ИИ – это основное, чему посвящена данная статья.

Благодарности

Авторы благодарны сотрудникам лаборатории Открытых информационных технологий кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова за обсуждения и ценные замечания.

Статья продолжает серию публикаций, начатых работой об обосновании исследований, посвященных устойчивым моделям машинного обучения [33]. Традиционно отмечаем, что все публикации в журнале INJOIT, связанные с цифровой повесткой, начинались с работ В.П. Куприяновского и его многочисленных соавторов [34-36].

Библиография

[1] Cyber Risk https://www.theirm.org/what-we-say/thought-leadership/cyber-risk/ Retrieved: Sep, 2024.
[2] Что такое киберриски и как застраховать свой бизнес https://ir.alfastrah.ru/posts/271 . Retrieved: Sep, 2024.
[3] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Искусственный интеллект и кибербезопасность." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[4] Намиот, Д. Е., and Е. А. Ильюшин. "Порождающие модели в машинном обучении." International Journal of Open Information Technologies 10.7 (2022): 101-118.
[5] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15.3 (2024): 1-45.
[6] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
[7] Mudarova, Ramina, and Dmitry Namiot. "Countering Prompt Injection attacks on large language models." International Journal of Open Information Technologies 12.5 (2024): 39-48. (in Russian)
[8] Namiot, Dmitry, and Elena Zubareva. "About AI Red Team." International Journal of Open Information Technologies 11.10 (2023): 130-139.
[9] Wach, Krzysztof, et al. "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT." Entrepreneurial Business and Economics Review 11.2 (2023): 7-30.
[10] Eiras, Francisco, et al. "Risks and Opportunities of Open-Source Generative AI." arXiv preprint arXiv:2405.08597 (2024).
[11] Duffourc, Mindy, and Sara Gerke. "Generative AI in health care and liability risks for physicians and safety concerns for patients." Jama (2023).
[12] 14 Risks and Dangers of Artificial Intelligence (AI) https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence Retrieved: 11.09.2024.
[13] Namiot, Dmitry. "On cyberattacks using Artificial Intelligence systems." International Journal of Open Information Technologies 12.9 (2024): 132-141. (in Russian)
[14] OWASP Top 10 for Large Language Model Applications https://owasp.org/www-project-top-10-for-large-language-model-applications/ Retrieved: 11.09.2024.
[15] AI Risk Repository https://airisk.mit.edu/ Проверено 11.09.2024
[16] NIST Trustworthy and Responsible AI - 600-1 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf Проверено 11.09.2024
[17] Pathmanathan, Pankayaraj, et al. "Is poisoning a real threat to LLM alignment? Maybe more so than you think." arXiv preprint arXiv:2406.12091 (2024).
[18] Bowen, Dillon, et al. "Scaling Laws for Data Poisoning in LLMs." arXiv preprint arXiv:2408.02946 (2024).
[19] Xu, Zihao, et al. "LLM Jailbreak Attack versus Defense Techniques--A Comprehensive Study." arXiv preprint arXiv:2402.13457 (2024).
[20] Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." arXiv preprint arXiv:2306.05499 (2023).
[21] Galli, Filippo, Luca Melis, and Tommaso Cucinotta. "Noisy Neighbors: Efficient membership inference attacks against LLMs." arXiv preprint arXiv:2406.16565 (2024).
[22] Maini, Pratyush, et al. "LLM Dataset Inference: Did you train on my dataset?." arXiv preprint arXiv:2406.06443 (2024).
[23] LLM AI Cybersecurity & Governance Checklist https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.1.pdf Проверено 11.09.2024
[24] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[25] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.
[26] AI RISK MANAGEMENT FRAMEWORK https://www.nist.gov/itl/ai-risk-management-framework Проверено 11.09.2024
[27] Research Initiative: AI Red Teaming & Evaluation https://genai.owasp.org/2024/09/12/research-initiative-ai-red-teaming-evaluation/ Проверено 11.09.2024
[28] OWASP Top 10 for LLM AI Red Teaming Methodologies, Guidelines, and Best Practices https://docs.google.com/document/d/1m06DMhonGuq8hTN30S-fAsuBA-ZK1UHMyzZamsZSTaE/edit Проверено 11.09.2024
[29] MIT Researchers Create an AI Risk Repository https://ide.mit.edu/insights/mit-researchers-create-an-open-ai-risk-repository/ Проверено 11.09.2024
[30] Slattery, Peter, et al. "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence." arXiv preprint arXiv:2408.12622 (2024).
[31] International Scientific Report on the Safety of Advanced AI https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
[32] Verma, Apurv, et al. "Operationalizing a Threat Model for Red-Teaming Large Language Models (LLMs)." arXiv preprint arXiv:2407.14937 (2024).
[33] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Текущие академические и индустриальные проекты, посвященные устойчивому машинному обучению." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[34] Цифровая экономика = модели данных + большие данные + архитектура + приложения? / В. П. Куприяновский, Н. А. Уткин, Д. Е. Намиот, П. В. Куприяновский // International Journal of Open Information Technologies. – 2016. – Т. 4, № 5. – С. 1-13. – EDN VWANDZ.
[35] Развитие транспортно-логистических отраслей Европейского Союза: открытый BIM, Интернет Вещей и кибер-физические системы / В. П. Куприяновский, В. В. Аленьков, А. В. Степаненко [и др.] // International Journal of Open Information Technologies. – 2018. – Т. 6, № 2. – С. 54-100. – EDN YNIRFG.
[36] Умная инфраструктура, физические и информационные активы, Smart Cities, BIM, GIS и IoT / В. П. Куприяновский, В. В. Аленьков, И. А. Соколов [и др.] // International Journal of Open Information Technologies. – 2017. – Т. 5, № 10. – С. 55-86. – EDN ZISODV.

On Cyber Risks of Generative Artificial Intelligence

Dmitry Namiot, Eugene Ilyushin

***Abstract*— This article is devoted to an overview of the risks of generative models of Artificial Intelligence. The rapid development of large language models has seriously increased attention to the security of Artificial Intelligence models. From a practical point of view, in this case, we are talking about the security of deep learning models. Large language models are susceptible to poisoning attacks, evasion attacks, attacks aimed at extracting training data, etc. But, at the same time, new attacks appear that are related specifically to the generated content. Moreover, the latter constitute an obvious majority. Therefore, recently many works have appeared that try to systematize all the risks of generative models. For example, OWASP and NIST are engaged in this. A complete taxonomy of generative AI risks should serve as a basis for building testing systems for generative models. This paper provides an overview of generative AI risk specifications outlined by OWASP, the NIST profile, and the MIT risk repository. The purpose of such specifications is to create a base for testing generative models and tools intended for AI Red Teams.**

***Keywords*— cyber risks, LLM, generative models.**

REFERENCES

[1] Cyber Risk https://www.theirm.org/what-we-say/thought-leadership/cyber-risk/ Retrieved: Sep, 2024.
[2] Chto takoe kiberriski i kak zastrahovat' svoj biznes https://ir.alfastrah.ru/posts/271 . Retrieved: Sep, 2024.
[3] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Iskusstvennyj intellekt i kiberbezopasnost'." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[4] Namiot, D. E., and E. A. Il'jushin. "Porozhdajushhie modeli v mashinnom obuchenii." International Journal of Open Information Technologies 10.7 (2022): 101-118.
[5] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15.3 (2024): 1-45.
[6] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
[7] Mudarova, Ramina, and Dmitry Namiot. "Countering Prompt Injection attacks on large language models." International Journal of Open Information Technologies 12.5 (2024): 39-48. (in Russian)
[8] Namiot, Dmitry, and Elena Zubareva. "About AI Red Team." International Journal of Open Information Technologies 11.10 (2023): 130-139.
[9] Wach, Krzysztof, et al. "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT." Entrepreneurial Business and Economics Review 11.2 (2023): 7-30.
[10] Eiras, Francisco, et al. "Risks and Opportunities of Open-Source Generative AI." arXiv preprint arXiv:2405.08597 (2024).
[11] Duffourc, Mindy, and Sara Gerke. "Generative AI in health care and liability risks for physicians and safety concerns for patients." Jama (2023).
[12] 14 Risks and Dangers of Artificial Intelligence (AI) https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence Retrieved: 11.09.2024.
[13] Namiot, Dmitry. "On cyberattacks using Artificial Intelligence systems." International Journal of Open Information Technologies 12.9 (2024): 132-141. (in Russian)
[14] OWASP Top 10 for Large Language Model Applications https://owasp.org/www-project-top-10-for-large-language-model-applications/ Retrieved: 11.09.2024.
[15] AI Risk Repository https://airisk.mit.edu/ Provereno 11.09.2024
[16] NIST Trustworthy and Responsible AI - 600-1 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf Provereno 11.09.2024
[17] Pathmanathan, Pankayaraj, et al. "Is poisoning a real threat to LLM alignment? Maybe more so than you think." arXiv preprint arXiv:2406.12091 (2024).
[18] Bowen, Dillon, et al. "Scaling Laws for Data Poisoning in LLMs." arXiv preprint arXiv:2408.02946 (2024).
[19] Xu, Zihao, et al. "LLM Jailbreak Attack versus Defense Techniques--A Comprehensive Study." arXiv preprint arXiv:2402.13457 (2024).
[20] Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." arXiv preprint arXiv:2306.05499 (2023).
[21] Galli, Filippo, Luca Melis, and Tommaso Cucinotta. "Noisy Neighbors: Efficient membership inference attacks against LLMs." arXiv preprint arXiv:2406.16565 (2024).
[22] Maini, Pratyush, et al. "LLM Dataset Inference: Did you train on my dataset?." arXiv preprint arXiv:2406.06443 (2024).
[23] LLM AI Cybersecurity & Governance Checklist https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.1.pdf Provereno 11.09.2024
[24] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[25] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.
[26] AI RISK MANAGEMENT FRAMEWORK https://www.nist.gov/itl/ai-risk-management-framework Provereno 11.09.2024
[27] Research Initiative: AI Red Teaming & Evaluation https://genai.owasp.org/2024/09/12/research-initiative-ai-red-teaming-evaluation/ Provereno 11.09.2024
[28] OWASP Top 10 for LLM AI Red Teaming Methodologies, Guidelines, and Best Practices https://docs.google.com/document/d/1m06DMhonGuq8hTN30S-fAsuBA-ZK1UHMyzZamsZSTaE/edit Provereno 11.09.2024
[29] MIT Researchers Create an AI Risk Repository https://ide.mit.edu/insights/mit-researchers-create-an-open-ai-risk-repository/ Provereno 11.09.2024
[30] Slattery, Peter, et al. "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence." arXiv preprint arXiv:2408.12622 (2024).
[31] International Scientific Report on the Safety of Advanced AI https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
[32] Verma, Apurv, et al. "Operationalizing a Threat Model for Red-Teaming Large Language Models (LLMs)." arXiv preprint arXiv:2407.14937 (2024).
[33] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Tekushhie akademicheskie i industrial'nye proekty, posvjashhennye ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[34] Cifrovaja jekonomika = modeli dannyh + bol'shie dannye + arhitektura + prilozhenija? / V. P. Kuprijanovskij, N. A. Utkin, D. E. Namiot, P. V. Kuprijanovskij // International Journal of Open Information Technologies. – 2016. – T. 4, # 5. – S. 1-13. – EDN VWANDZ.

[35] Razvitie transportno-logisticheskih otraslej Evropejskogo Sojuza: otkrytyj BIM, Internet Veshhej i kiber-fizicheskie sistemy / V. P. Kuprijanovskij, V. V. Alen'kov, A. V. Stepanenko [i dr.] // International Journal of Open Information Technologies. – 2018. – T. 6, # 2. – S. 54-100. – EDN YNIRFG.

[36] Umnaja infrastruktura, fizicheskie i informacionnye aktivy, Smart Cities, BIM, GIS i IoT / V. P. Kuprijanovskij, V. V. Alen'kov, I. A. Sokolov [i dr.] // International Journal of Open Information Technologies. – 2017. – T. 5, # 10. – S. 55-86. – EDN ZISODV.