Некоторые аспекты внедрения основанных на искусственном интеллекте решений в комплексные системы безопасности объектов гражданского назначения

Косарев Владимир Евгеньевич
к.т.н., доцент, Финансовый университет при Правительстве Российской Федерации, vkosarev@fa.ru,

Смирнов Михаил Викторович
к.т.н., доцент, Финансовый университет при Правительстве Российской Федерации, mvsmirnov@fa.ru

Статья посвящена исследованию факторов, снижающих устойчивость функционирования нейросетевых алгоритмов в системах обеспечения комплексной безопасности объектов критической инфраструктуры. Авторы рассматривают проблему повышения уровня доверия к выходным решениям искусственного интеллекта (ИИ) через призму риск-ориентированного подхода и формализации диаграммы Исикавы. В работе проведена систематизация причин возникновения естественных ошибок глубоких нейронных сетей, связанных с пропуском целей, ошибочной классификацией объектов из-за пересечения таксонов, а также сжатием качества работы при изменении внешних условий и зашумлении фона. Особое внимание уделено угрозам вредоносного машинного обучения (Adversarial Machine Learning), включая атаки уклонения, спуфинг и отравление датасетов. Сделан вывод о необходимости проведения цифровых испытаний и разработки превентивных методов защиты ИИ-модулей.

Ключевые слова: искусственный интеллект, критическая инфраструктура, комплексная безопасность, риск-ориентированный подход, нейронные сети, вредоносное машинное обучение, доверие к ИИ.

Современные вызовы перед обществом и государством требуют внедрения новых информационных технологий в интегрированные комплексные системы безопасности. Качественно новые возможности для мониторинга внешней обстановки защищаемых объектов критической инфраструктуры открывают технологии искусственного интеллекта (ИИ). Внедрение технологий ИИ в контур интегрированных систем охраны и безопасности повышает защищенность объектов, снижает нагрузку на персонал подразделений защиты (безопасности), уменьшает субъективность в принятии решений. Применение нейронных сетей глубокого обучения значительно повышает автоматизацию управления в условиях многокритериальности и неочевидности. Однако принципы функционирования глубоких нейронных сетей не позволяют алгоритмическим путем убедиться в безупречности их работы [1].

Таким образом, внедрение технологий ИИ в комплексные системы безопасности объектов критической инфраструктуры требует особых усилий и применения новых подходов в целях повышения уровня доверия пользователя к выходному решению нейросетевых алгоритмов [2].

Для систематизации факторов, оказывающих наибольшее влияние на возможное снижение качества работы интеллектуальных модулей, представляется целесообразным применение методов риск-ориентированного анализа [3].

Основные риски и угрозы внедрения современных информационных технологий в контур технических средств обеспечения комплексной безопасности объектов критической инфраструктуры возможно представить в формализации диаграммы Исикавы (рисунок – 1).



Рисунок 1 – Риски внедрения технологий ИИ в системы комплексной безопасности объектов критической инфраструктуры

Факторный анализ теории и практики применения нейронных сетей, как основы современных систем ИИ, позволяет выделить группы рисков функционирования программного обеспечения (ПО), в частности:

1. *Отсутствие релевантной реакции* интеллектуальной системы обеспечения безопасности *на объект* оперативного интереса, очевидный для человеческого восприятия. Например, нередко при просмотре фотоснимка, снятого той же оптико-электронной системой (ОЭС) [4], оператор легко находит объект интереса, в то время как обученная нейросеть, казалось бы, в простых условиях применения, пропускает цель.

Примеры пропуска цели нередки также для онтологических, экспертных и языковых моделей, когда из текстовой информации легко определить интересующие элементы, но многоуровневая нейросетевая модель их пропускает. Причины таких явлений не всегда кроются в «банально» плохой обученности нейронной сети. Существуют внутренние причины «утыкания» градиентных способов обучения в локальный минимум и ошибочного выбора стратегии или траектории оптимизации в направлении псевдоградиента.

2. *Неверная классификация объекта интереса* или смещение собственных объектов и целей. Причины таких нарушений работы сети заключаются в пересечении классов (таксонов) при их классификации на стадии обучения. Нейронная сеть выстраивает синаптические связи многокритериально. Во многих случаях однозначно разделить классы для ИИ невозможно, возникают взаимопересеченные области (рисунок 2).

3. *Непредумышленные ошибки работы сети*, заключающиеся в том, что при незначительном изменении визуального образа объекта, происходит качественно (принципиально) неверная классификация объекта. Классическим примером является незначительная помарка на объекте идентификации,

приводящая к неудачному соотношению реакций синаптических связей и качественно неверному отклику сети. Вариантом таких ошибок могут быть блики на дорожном знаке, приводящие к восприятию запрещающего знака системой активной безопасности автомобиля как разрешающего знака рекомендованной скорости движения транспортного средства.



Рисунок 2 – Пример неоднозначной классификации при испытании сети

4. *Значительные ошибки нейросетевых систем при незначительных изменениях входных условий* функционирования. В частности, проблемы работы интеллектуальных модулей систем технического зрения могут значительно возрастать в ходе сезонных изменений фоноцелевой обстановки, времени суток применения интеллектуальных средств безопасности и неблагоприятных метеорологических условиях. Чаще, чем в иных случаях, ошибки классификации объекта оперативного интереса возникают в условиях недостаточной или ограниченной видимости в ночное время суток, при осадках и наличии факторов засветки объектива.

Ниже представлен пример ошибочного функционирования нейросетевых средств классификации и идентификации при незначительных внешних воздействиях (рисунок 3) для отечественного программно-аппаратного комплекса (ПАК), функционирующего в настоящее время на объектах критической гражданской инфраструктуры [5]. На данном примере камень телесного цвета, находящийся на противоположном от точки съемки берегу реки, идентифицируется нейронной сетью, как сидящий человек.



Рисунок 3 – Сбой при испытании отечественной ИИ-системы комплексной охраны объекта критической инфраструктуры

5. *Вредоносное воздействие на систему ИИ.* Развитие современных топологий свёрточных нейронных сетей, энкодеров и сетей-трансформеров сформировало базу для массового применения систем технического зрения и автоматической разметки данных широким кругом производителей зарубежных стран. Однако использование заимствованных (недоверенных) нейросетевых алгоритмов и использование однотипных подходов их применения приводит к легкому поиску злоумышленниками уязвимых мест таких систем. При этом, критические транзакции и точки бифуркации в процессах нейросетевого принятия решения трудно прогнозируемы.

Получающие все большее распространение методы вредоносного машинного обучения AML (Adversarial Machine Learning) несут в себе риски влияния на систему ИИ не только на стадии обучения, но и в процессе дообучения или функционирования сети (рисунок 4). Угрозу представляют атаки уклонения, нарушающие адекватность и предсказуемость работы сети, спуфинг-атаки, а также отравление дата-сета [5].

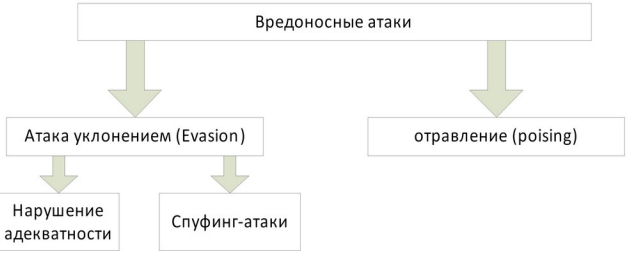


Рисунок 4 – Сбои в работе систем ИИ вследствие внешнего вредоносного возмущающего воздействия

Нейронные сети априори склонны к ошибкам классификации при наложении незначительных естественных возмущений на вход синапса нейрона. Применительно к техническим средствам охраны – идентификация лиц подвержена ошибкам 1-го рода. Примером является ошибка разблокировки устройств безопасности TouchID, несмотря на то, что отпечатки пальцев человека уникальны, а их текстура легко детерминируема и без применения технологий ИИ. Примером ошибок 2-го рода в системах контроля и управления доступом охраняемых объектов является нераспознание доверенного лица по ключевым точкам протоколов FaceID.

Атаки отравлением предполагают зашумление визуального образа внешним воздействием, которое с учетом специфики глубоких нейросетей может привести к ошибочной классификации объекта интереса. Таким воздействием может быть использование завес чёрного (белого) дымов, естественного и искусственного тумана, гари, пыли, песчаной бури, выхлопных газов дизельных автомобилей с исключенным сажевым фильтром, нагнетаемого снега. Наложение даже 0,5-процентной зернистости к исходному изображению, незаметное для подразделений охраны, как правило, приводит к неверному выходу автоматического нейросетевого классификатора объектов проникновения (рисунок 5).

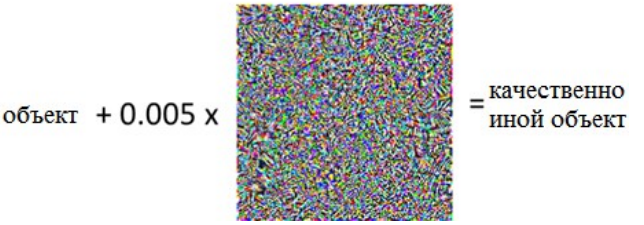


Рисунок 5 – Ошибка классификатора системы технического зрения ИИ при незначительном зашумлении основного образа

Одновременно с этим, снижения устойчивости функционирования нейронной сети FaceID-систем безопасности при знании специфики работы современных фреймворков потенциальному нарушителю легко добиться закрытием ключевых точек распознавания, кепкой или очками с кратковременной засветкой, накладной маской [8].



Рисунок 3.6 – Сбой системы ИИ при тестировании системы распознавания после воздействия атакой уклонения

Рассмотренные выше примеры естественных ошибок работы нейросетевых алгоритмов технических систем охраны и обеспечения безопасности объектов критической инфраструктуры подчеркивают необходимость более глубокого исследования их побочных эффектов как на стадии разработки, так и при проведении цифровых (виртуальных) испытаний современных интеллектуальных систем обеспечения комплексной безопасности.

Одновременно с этим, представленные подходы AML-вредоносного воздействия на системы ИИ требуют самого внимательного изучения и разработки соответствующих программ и методик испытаний, а по их итогам разработки дополнительных средств и способов обеспечения превентивной информационной защиты.

Литература

1. ГОСТ Р 59898-2021. Оценка качества систем искусственного интеллекта. Общие положения.

2. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения.
3. ГОСТ Р ИСО/МЭК 27005 Информационная технология. Методы и средства обеспечения безопасности. Менеджмент риска информационной безопасности.
4. Молчанов А.С., Коломоец В.А. Цифровые портреты типовых объектов воздушной разведки. – М.: Издательство «Перо», 2025 – 194 с.
5. Ермоленко А.В., Полушкин В.М., Смирнов М.В. Ошибки работы нейросетевых систем и вредоносное машинное обучение в задачах технического зрения. Журнал «Нейрокомпьютеры: разработка, применение». М: «Радиотехника», 2024, № 6 (ВАК) (стр. 23-30).
6. Fu J, Liu H, Tian H, Fang Z, Lu H. Dual attention network for scene segmentation. IEEE CVPR 2019: 3146-3154.
7. Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. IEEE Conf Comp Vis Pattern Recogn (CVPR) 2016: 770-778.
8. Lin G, Milan A, Shen Ch, Reid I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. IEEE CVPR 2019: 5168-5177.

Some Aspects of Implementing AI-Based Solutions in Integrated Security Systems for Civil Facilities

Kosarev V.E., Smirnov M.V.
Financial University under the Government of the Russian Federation

JEL classification: C01, C02, C1, C4, C5, C6, C8

This article examines factors that compromise the operational stability of neural network algorithms within integrated security systems for critical infrastructure facilities. The authors address the challenge of increasing trust in AI-generated outputs by applying a risk-based approach and formalizing an Ishikawa diagram. The study systematizes the causes of inherent errors in deep neural networks—such as missed targets, object misclassification due to overlapping taxonomic categories, and performance degradation caused by changing environmental conditions or background noise. Particular attention is paid to threats associated with adversarial machine learning, including evasion attacks, spoofing, and dataset poisoning. The authors conclude that there is a need for digital testing and the development of preventive protection methods for AI modules.

Keywords: artificial intelligence, critical infrastructure, integrated security, risk-based approach, neural networks, adversarial machine learning, trust in AI.

References

1. GOST R 59898-2021. Artificial intelligence systems quality assessment. General.
2. GOST R 59276-2020. Artificial intelligence systems. Methods for ensuring trust. General.
3. GOST R ISO/IEC 27005. Information technology. Security techniques. Information security risk management.
4. Molchanov A.S., Kolomoets V.A. Tsifrovye portrety tipovykh obektov vozdushnoy razvedki. – M.: Izdatelstvo «Pero», 2025 – 194 pp. (In Russian).
5. Ermolenko A.V., Polushkin V.M., Smirnov M.V. Errors of Neural Network Systems and Adversarial Machine Learning in Computer Vision Tasks. Neirokomp'iutery: razrabotka, primenenie [Neurocomputers: Development, Application]. Moscow: Radiotekhnika, 2024, no. 6, pp. 23-30. (In Russian).
6. Fu J, Liu H, Tian H, Fang Z, Lu H. Dual attention network for scene segmentation. IEEE CVPR 2019: 3146-3154.
7. Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. IEEE Conf Comp Vis Pattern Recogn (CVPR) 2016: 770-778.
8. Lin G, Milan A, Shen Ch, Reid I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. IEEE CVPR 2019: 5168-5177.