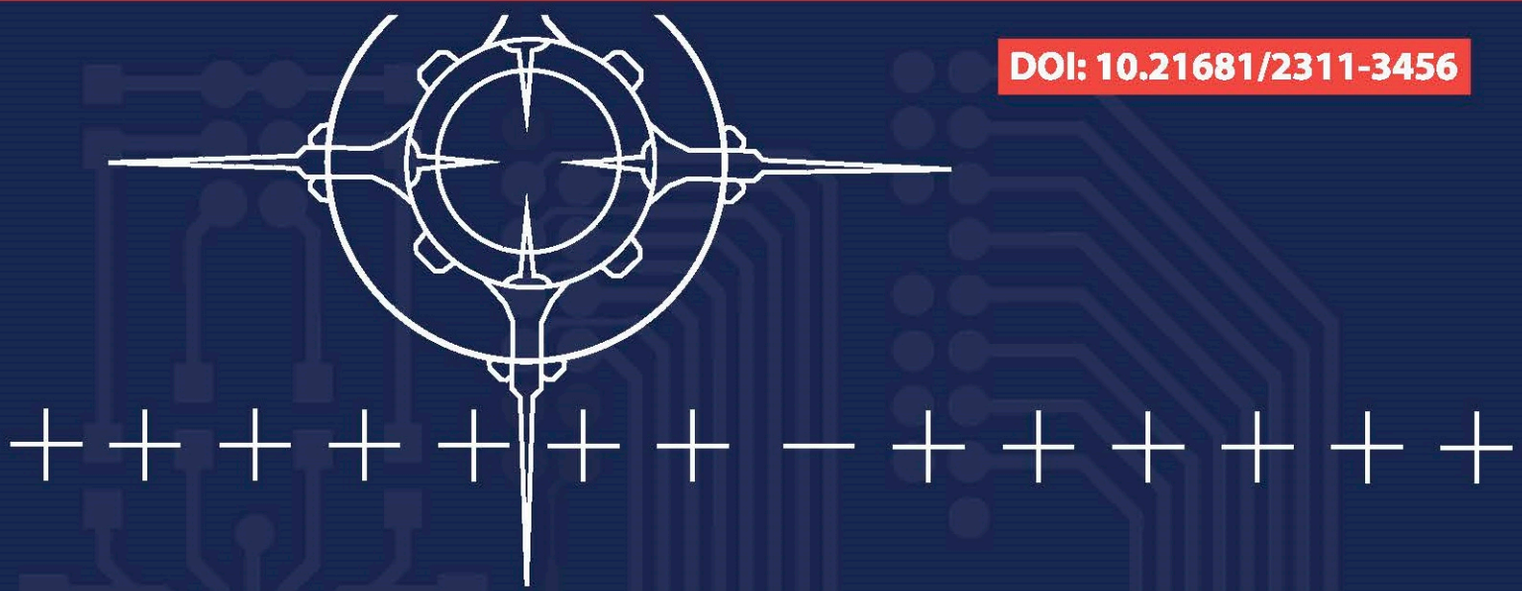
ВОПРОСЫ КИБЕРБЕЗОПАСНОСТИ

№1 2024 (59)

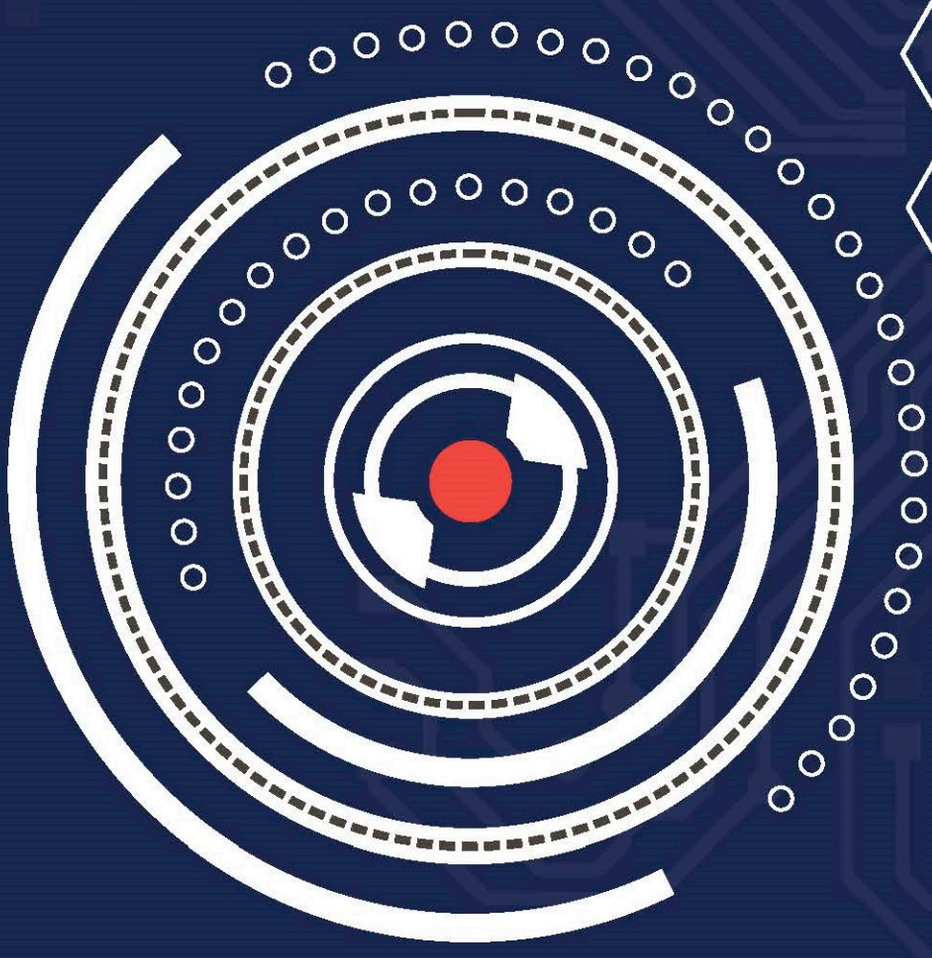
DOI: 10.21681/2311-3456

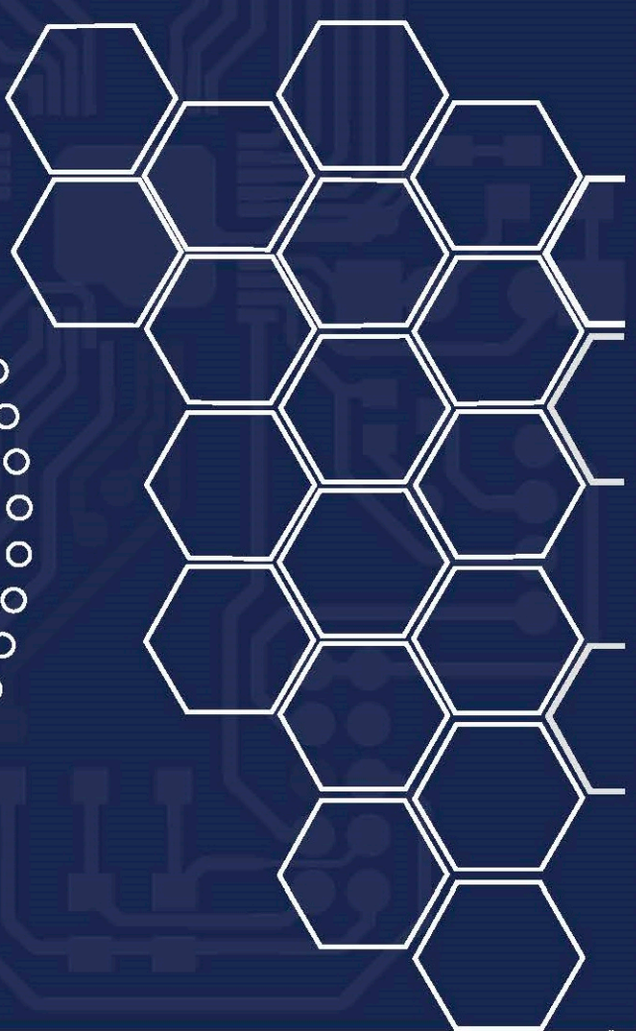


Тренды и модели технологий искусственного интеллекта

Методы сетевой безопасности

Риски безопасности микросетей





Эшелон
комплексная безопасность

{KOMRAD}
Enterprise SIEM

ВЫСОКАЯ ПРОИЗВОДИТЕЛЬНОСТЬ И МИНИМАЛЬНЫЕ ТРЕБОВАНИЯ К АППАРАТНОМУ ОБЕСПЕЧЕНИЮ

СЗИ СЗИ

KOMRAD Enterprise SIEM позволяет осуществлять централизованный сбор событий ИБ, выявлять инциденты ИБ и оперативно на них реагировать. Применение комплекса позволяет эффективно выполнять требования, предъявляемые регуляторами к защите персональных данных, к обеспечению безопасности государственных информационных систем и контролю критической информационной инфраструктуры предприятия. KOMRAD позволяет отправлять данные о событиях и инцидентах ИБ во внешние системы (например, ГосСОПКА).

- Визуальный конструктор запросов и директив корреляции
- Высокая производительность
- Гибкая интеграция с нестандартными источкниками событий
- Широкий спектр поддержки источников событий
- Ролевая модель управления доступом
- Оперативное оповещение об инциденте
- Масштабируемость

Чтобы получить демо-версию KOMRAD Enterprise SIEM или заказать пилот у наших партнеров в вашем регионе, свяжитесь с нашим отделом продаж по e-mail: sales@npo-echelon.ru.

ВОПРОСЫ КИБЕРБЕЗОПАСНОСТИ

НАУЧНЫЙ РЕЦЕНЗИРУЕМЫЙ ЖУРНАЛ

№1 (59) 2024 г.

Выходит 6 раз в год

Журнал выходит с 2013 г. (Свидетельство о регистрации ПИ № ФС77-75239). Перерегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 07.03.2019.

Журнал входит в рейтинг научных изданий ВАК в категории К1, а также в российский индекс научного цитирования RSCI на международной платформе Web of Science (WoS)

Учредитель и издатель

АО «Научно-производственное объединение «Эшелон»

Над номером работали:

Г. И. Макаренко – *шеф-редактор*, И. Ю. Шахалов – *отв. секретарь*, Т. В. Галатонов – *сайт*, Н. С. Рождественская – *маркетинг и подписка*

Подписано к печати 20.01.2024 г.
Общий тираж 120 экз. Цена свободная

Адрес: 107023, Москва, ул. Электрозаводская, д. 24, стр. 1.
E-mail: editor@cyberrus.info, тел.: +7 (985) 939-75-01.
Требования, предъявляемые к рукописям, размещены на сайте: **https://cyberrus.info/**

СОДЕРЖАНИЕ

Подписка на журнал осуществляется в почтовых отделениях по каталогу «Пресса России». Подписной индекс 40707

ПРАВОВЫЕ ГОРИЗОНТЫ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: НАЦИОНАЛЬНЫЙ И МЕЖДУНАРОДНЫЙ АСПЕКТ.

Карцхия А. А.[1], Макаренко Г. И.[2]

DOI: 10.21681/2311-3456-2024-1-2-14

Цель исследования *– анализ факторов стремительного развития искусственного интеллекта и его потенциала, исследование новых моделей ИИ для повышения производительности труда, поощрения инноваций и формирования новых предпринимательских структур, а также решения социальных проблем в здравоохранении, образовании, разрешении климатического кризиса и достижении целей устойчивого развития.*

Методы исследования: *сравнительно правовой метод и методы анализа и синтеза в процессе исторического генезиса искусственного интеллекта, применение риск ориентированного метода оценки ИИ.*

Результат: *в исследование показано, что внедрение программ искусственного интеллекта вместе с преимуществами создает трудно прогнозируемые угрозы и риски, имеющие трансграничный характер. С целью смягчения потенциальных опасностей, для обеспечения контролируемости и устойчивости технологий ИИ на основе концепции доверенного (ответственного) искусственного интеллекта необходимо утверждение руководящих принципов по искусственному интеллекту и создания универсального кодекса поведения разработчиков ИИ, которые совместно могут создать базу для единых основ правового регулирования в рамках национального законодательства каждой страны на основе принципов защиты прав человека, конфиденциальности и защиты данных, а также прозрачности и объяснимости, справедливости, подотчетности и безопасности ИИ, надлежащего контроля со стороны человека и этических норм создания и применения ИИ.*

Новизна исследования *заключается в том, что на основе риск ориентированного подхода предлагается концептуальная оценка полезности эффективности, устойчивости и безопасности технологий и моделей ИИ, а также установления его правового статуса, в том числе, для защиты человека от неконтролируемого влияния ИИ и неизменности гарантий конституционных прав и свобод человека.*

Ключевые слова: *международная кибербезопасность, доверенный интеллект, нейронные сети, машинное обучение, безопасность искусственного интеллекта, технологический суверенитет, угрозы и риски технологий, устойчивое развитие.*

LEGAL HORIZONS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES: NATIONAL AND INTERNATIONAL ASPECTS.

Kartskhia A. A.[3], Makarenko G. I.[4]

The purpose *of the study is to analyze the factors of rapid development of artificial intelligence and its potential, to study new AI models to increase labor productivity, encourage innovation and the formation of new business structures, as well as solve social problems in healthcare, education, solving the climate crisis and achieving sustainable development goals.*

Research methods: *comparative legal method and methods of analysis and synthesis in the process of historical genesis of artificial intelligence, application of the risk-oriented method of AI assessment.*

Result: *The study shows that the introduction of artificial intelligence programs, together with the benefits, creates hard-to-predict threats and risks of a cross-border nature. In order to mitigate potential dangers, in order to ensure the controllability and sustainability of AI technologies based on the concept of trusted (responsible) artificial intelligence, it is necessary to approve guidelines on artificial intelligence and create a universal code Together, AI can create a framework for a common regulatory framework in each country's national legislation based on the principles of human rights, privacy and data protection, as well as transparency and explainability, fairness, accountability and safety of AI, appropriate human control, and ethical standards for the creation and use of AI.*

1 Карцхия Александр Амиранович, доктор юридических наук, профессор РГУ нефти и газа (НИУ) имени И. М. Губкина, Москва. Россия. E-mail: arhz50@mail.ru
2 Макаренко Григорий Иванович, старший научный сотрудник НЦПИ при Минюсте РФ, Москва, Россия. E mail: t7920518@yandex.com

3 Alexander A. Kartskhia, Doctor of Law, Professor, Gubkin Russian State University of Oil and Gas, Moscow. Russia. E-mail: arhz50@mail.ru
4 Grigory I. Makarenko, Senior Researcher, National Center for Strategic Studies under the Ministry of Justice of the Russian Federation, Moscow, Russia. E-mail: t7920518@yandex.com

The novelty of the research *lies in the fact that, on the basis of a risk-oriented approach, a conceptual assessment of the usefulness of the efficiency, sustainability and safety of AI technologies and models, as well as the establishment of its legal status, including for the protection of a person from the uncontrolled influence of AI and the immutability of guarantees of constitutional human rights and freedoms, is proposed.*

Keywords: *international cybersecurity, trusted intelligence, neural networks, machine learning, artificial intelligence security, technological sovereignty, threats and risks of technologies, sustainable development.*

Введение. Искусственный интеллект в центре всеобщего внимания

Мир стоит на пороге нового ренессанса в науке и технике, основанного на всестороннем понимании структуры и поведения материи от наноразмерных величин до самой сложной из когда-либо открытых систем - человеческого мозга. При должном внимании к этическим вопросам и потребностям общества результатом может стать значительное улучшение человеческих способностей, новых отраслей промышленности и продуктов, результатов для общества и качества жизни. При этом, все более актуальными становятся вопросы правового регулирования этих сфер деятельности [1,2,3][5].

Как отмечается в Концепции внешней политики Российской Федерации[6], человечество переживает эпоху революционных перемен. Структурная перестройка мировой экономики, связанная с переходом на новую технологическую основу посредством внедрения технологий искусственного интеллекта, новейших информационно-коммуникационных, энергетических, биологических технологий и нанотехнологий, а также рост национального самосознания, культурно-цивилизационное разнообразие и другие объективные факторы ускоряют процессы перераспределения потенциала развития в пользу новых центров экономического роста и геополитического влияния.

Экспоненциальные улучшения технологий искусственного интеллекта и других передовых технологий в последнее время привели к взрывному росту интереса (научного, коммерческого, военного и др.) в искусственный интеллект и финансовых инвестиций в него.

В настоящее время особое внимание приковано к *генеративному искусственному интелекту*. Генеративный искусственный интеллект (*AGI*) – это тип искусственного интеллекта, который может создавать (генерировать) новый контент и идеи, включая разговоры, истории, изображения, видео и музыку. Как и любой искусственный интеллект, генеративный ИИ основан на моделях машинного обучения – очень больших моделях, предварительно обученных на **огромных** объемах данных и обычно называемых базовыми моделями (FM). Базовые модели (FM), обученные работе с огромными наборами данных, представляют собой крупные нейронные сети с глубоким обучением, которые изменили подход специалистов по работе с данными к машинному обучению (ML). Вместо того чтобы разрабатывать искусственный интеллект с нуля, специалисты по работе с данными используют базовую модель в качестве отправной точки для разработки моделей ML, позволяющих быстрее и экономичнее поддерживать новые сферы применения. Термин «*базовая модель*» был придуман исследователями для описания моделей ML, обученных на широком спектре обобщенных и немаркированных данных и способных выполнять широкий спектр общих задач, таких как понимание языка, генерирование текста и изображений и общение на естественном языке. Технологии искусственного интеллекта пытаются имитировать человеческий интеллект в таких нетрадиционных вычислительных задачах, как распознавание изображений, обработка естественного языка (NLP) и перевод. Генеративный искусственный интеллект является следующим шагом в разработке искусственного интеллекта[7].

Генеративный ИИ быстро вошел в общественный дискурс. Стремительный прогресс в области генеративного искусственного интеллекта обусловлен его ожидаемым потенциалом для повышения производительности, поощрение инноваций и предпринимательства и поиск решений глобальных проблем, а также в решении социальных проблем, таких как улучшение здравоохранения и помощь в разрешении климатического кризиса и достижении Целей устойчивого развития (ЦУР).

5 Roco, M.C., Bainbridge, W.S. (2003). Overview Converging Technologies for Improving Human Performance. In: Roco, M.C., Bainbridge, W.S. (eds) Converging Technologies for Improving Human Performance. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-0359-8_1

6 Указ Президента РФ от 31.03.2023 N 229 «Об утверждении Концепции внешней политики Российской Федерации»// Собрание законодательства РФ, 03.04.2023, N 14, ст. 2406

7 https://aw.club/global/ru/blog/ai/generative-ai-for-content-creation (дата обращения 11 января 2024 г.)

Современные понятия искусственного интеллекта различаются в своих определениях. Некоторые фокусируются на способности программы делать любые прогнозы, рекомендации или решения. Например, NIST (*NIST Artificial Intelligence Risk Management Framework*) определяет «систему ИИ» в широком смысле как любую «машинную систему, которая может для заданного набора целей генерировать результаты, такие как прогнозы, рекомендации или решения, влияющие на реальную или виртуальную среду»[8]. Другие определения дают более узкое определение термина для обозначения программ, которые либо приближают человеческое мышление, либо заменяют его, т.е. программы, которые способны приближаться к человеческим, интеллектуальным способностям.

Искусственный интеллект рассматривается в российской Национальной стратегии развития искусственного интеллекта и российском законодательстве[9] как комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Такой комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру (в том числе информационные системы, информационно-телекоммуникационные сети, иные технические средства обработки информации), программное обеспечение (в т.ч. в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

При этом следует учесть, что технологии ИИ классифицируются в трех разных аспектах: методы, применяемые при создании ИИ (например, машинное обучение); функциональные приложения (например, обработка речи и компьютерное зрение); и области применения этих технологий (например, связь, транспорт)[10].

Стоит отметить, что термин «искусственный интеллект», как отмечают эксперты [4,5][11], не относится к какой-либо конкретной технологии – скорее, это собирательный термин для множества технологий использования математико-статистических методов для моделирования когнитивных способностей. Технологии искусственного интеллекта работают на основе анализа большого объема неструктурированных данных (*Big Data*) по специально разработанному алгоритму для выявления определенных закономерности данных и получения на их основе конкретного вывода с использованием нейронной сети, алгоритмы и структура которых основаны на функциональных принципах человеческого мозга, где большое количество отдельных алгоритмов работают вместе во взаимосвязанном и взаимозависимым образом, отражающим функционирование сети синапсов в человеческом мозге.

Сложные нейронные сети с несколькими уровнями обработки (со множеством соединенных последовательно и влияющих друг на друга алгоритмов) называются ***глубокими*** *нейронными сетями* (*Deep Neural Networks*). В сложных («глубоких») нейронных сетях способ взаимодействия отдельных алгоритмов друг с другом больше не определяется разработчиком, поскольку количество определяемых параметров слишком велико. Вместо этого подходящие обучающие данные (т. е. обучающие данные, специально отобранные и предназначенные для использования по назначению) передаются в нейронную сеть для обработки в автоматических циклах обучения. Нейронная сеть использует процессы статистической оптимизации для определения наиболее подходящих настроек (параметризация), например, для автономной идентификации лица на снимках. Этот процесс автоматической параметризации нейронной сети известен как глубокое обучение (*Deep Learning*). Качественный уровень технологии ИИ зависит от его архитектуры, обучения и качества обучающих данных, поскольку структура нейронной сети, ее настройки должны быть адаптированы к конкретной цели, на которую нацелен ИИ (например, распознавание речи или изображений, генерация текста и т.д.). В идеале приложение искусственного интеллекта должно быть способно идентифицировать в большом объеме данных (например, в потоке данных камеры наблюдения) тип шаблона, для которого оно было обучено (например, лица, номерные знаки и т. д.), за очень короткое время. Фактический показатель успешности приложений искусственного интеллекта во многом зависит от структуры нейронной сети, способа ее обучения и качества используемых обучающих данных.

В связи с этим, важно установить **современное понимание искусственного интеллекта**. Обычно, *генеративный (общий) искусственный интеллект* (*AI*)

8 Shukla Shubhendu et al, Applicability of Artificial Intelligence in Different Fields of Life, 1 Int'l J. of Scientific Engineering and Research 1 at 28 (Sept. 2013).

9 Указ Президента РФ от 10.10.2019 N 490 «О развитии искусственного интеллекта в Российской Федерации // Собрание законодательства РФ, 14.10.2019, N 41, ст. 5700; Федеральный закон от 24.04.2020 N 123-ФЗ// Собрание законодательства РФ, 27.04.2020, N 17, ст. 2701

10 WIPO Technology Trends 2019, Artificial Intelligence. URL: https://www.theblockchaintest.com/uploads/resources/WIPO%20-%20Technology%20Trends%202019-Artificial%20Intelligence%20-%202019.pdf

11 Russell S., Norvig P. Artificial Intelligence: A Modern Approach. Third Edition. Boston: Prentice Hall, 2010. xviii; 1132 p. P. 1–2;
Rissland E.L. Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning // The Yale Law Journal. 1990. Vol. 99. N 8. P. 1957-1981. P. 1958-1959.

на основе машинного обучения использует нейронные сети и другие алгоритмы для создания новых данных или контента, похожих на исходные данные. Этот подход отличается от дескриптивного AI, который анализирует и классифицирует данные, но не создает новых данных. Генеративный AI может иметь огромное значение для различных отраслей, таких как медиа, искусство, развлечения, реклама и образование. Однако, он также может вызывать определенные угрозы в связи с нарушением авторских прав, распространением ложной или дискриминационной информации и потери контроля над созданным контентом.

Дескриптивный искусственный интеллект (AI) на основе машинного обучения используется для анализа, классификации и предсказания на основе необработанных данных и определяет структуру, зависимости и тенденции данных, не создавая новых данных. Дескриптивный AI может быть использован для различных целей, таких как: (a) классификация, т.е. разделение данных на группы на основе их характеристик или признаков (классификация электрокардиограмм (ЭКГ) на нормальные и аномальные, диагностика заболеваний и др.); (b) регрессия, т.е. предсказание неизвестных значений на основе известных данных (прогноз погоды, биржевых котировок и др.); (c) кластеризация, т.е. разделение данных на группы на основе схожести между элементами (моделирование бизнес-процессов и др.); (d) анализ тенденций, т.е. определение тенденций и зависимостей в данных для получения информации о будущих событиях или изменениях. Дескриптивный AI является основой для многих современных технологий, таких как рекомендательные системы, системами автоматической обработки звука и изображений, системами контроля качества и системами управления рисками. Хотя дескриптивный AI не создает новых данных, он может предоставить важную информацию и знания, которые могут быть использованы для принятия решений, планирования и стратегического планирования.

Вместе с тем, уже разработан новый вид ИИ – ***саморазвивающийся*** *искусственный интеллект*. Как заявили ученые Массачусетского технологического института и Калифорнийского университета (Fox News)[12], возможно создание подсистем ИИ без помощи человека. Более крупные модели ИИ, подобные тем, которые используют ChatGPT, на основе «родительского» алгоритма могут создавать меньшие, специфичные приложения искусственного интеллекта, которые можно применять, например, для усовершенствования слуховых аппаратов, мониторинга нефтепроводов или отслеживания исчезающих видов живой природы.

Ведущие страны в сфере разработки искусственного интеллекта, опираясь на активную поддержку государства, стремительными темпами развивают национальные технологии ИИ. После разработок Deep Mind и запуска в ноябре 2022 американской Open AI *ChatGPT* последовали публичные старты аналогичных технологий на основе LLM в других странах. В ноябре 2023 года в Абу-Даби (ОАЭ) была запущена поддерживаемая государством компания по искусственному интеллекту AI71 для коммерциализации модели ИИ LLM Falcon. В декабре того же года объявлено о масштабном финансировании французского ИИ Mistral. В Индии создаются национальные модели LLM Krutrim и Sarvam. Государства и частные компании в США, КНР, Великобритании, Франции, Германии, Индии, Саудовской Аравии и Объединенных Арабских Эмиратах (ОАЭ) масштабно финансируют разработки ИИ и развивают национальные производства графических редакторов (GPU-чипов) и других элементов, необходимых для создания ИИ[13]. В России созданы аналогичные разработки и достижения на базе нейросети ПАО Сбербанк России (RuGPT-3).

Международно-правовые аспекты статуса искусственного интеллекта

На первом международном саммите по безопасности искусственного интеллекта, прошедшем 1 ноября 2023г. в Великобритании (Россия не принимала в нем участие), страны-участники, включая США, Европейский Союз, Великобританию, Францию, Германию, Италию, КНР, Австралию Индию, Бразилию, Японию, Королевство Саудовской Аравии, Объединенные Арабские Эмираты, Нигерию и Кению, а также компании-лидеры IT отрасли (Amazon Web Services, Anthropic, Google, Google DeepMind, Inflection AI, Microsoft, Mistral AI, Open AI и xAI) подписали Декларацию по вопросам безопасности искусственного интеллекта (*The Bletchley Declaration on AI safety*)[14] (далее – Декларация), в которой устанавливается, что искусственный интеллект открывает огромные глобальные возможности, обладая потенциалом трансформировать мир и повышать благосостояние людей, и потому, для всеобщего блага искусственный интеллект должен проектироваться, разрабатываться, развертываться и использоваться безопасным образом.

12 Ученые заявили о возможности ИИ воспроизводиться без участия человека, 17 декабря 2023. URL: https://vfokuse.mail.ru/article/uchenye-zayavili-o-vozmozhnosti-ii-vosproizvoditsya-bez-uchastiya-cheloveka-59040575/ (дата обращения 11.01.2024 г.)

13 Welcome to the era of AI nationalism. The Economist, January 1st, 2024. URL: https://www.economist.com/ business/2024/01/01/welcome-to-the-era-of-ai-nationalism?utm_content=article-link-2&etear=nl_today_2&utm_campaign=a.the-economist-today&utm_medium=email.internal-newsletter.np&utm_source=salesforce-marketing-cloud&utm_term=1/1/2024&utm_id=1840347

14 https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration

Подтверждается необходимость безопасного развития искусственного интеллекта и использования его преобразующих возможностей во благо всех, инклюзивным образом во всем мире, включая сферу здравоохранения и образования, продовольственной безопасности, науки, чистой энергетики, биоразнообразия и климата, а также для реализации прав человека и активизации усилий по достижению Целей устойчивого развития ООН. Однако огромные возможности ИИ сопряжены с рисками, которые могут угрожать глобальной стабильности.

В Декларации использован термин «*Frontier AI*» (*передовой, новаторский ИИ*), который представляет собой высокоэффективные модели ИИ общего назначения, которые могут выполнять широкий спектр задач и соответствовать или превосходить возможности, присутствующие в самых продвинутых моделях на сегодняшний день. В первую очередь это относится к большим языковым моделям (LLM), лежащим в основе ChatGPT,4 Claude,5 и Bard.6. Ведущие компании в области ИИ, такие как Open AI, DeepMind и Anthropic, разрабатывают большие языковые модели (LMS), такие как GPT-4, в два этапа: предварительное обучение и тонкая настройка. На предварительном этапе обучения LLM «читает» миллионы или миллиарды текстовых документов, обучаясь выстраивать слова. Во время тонкой настройки предварительно обученный ИИ дополнительно обучается на тщательно отобранных наборах данных, которые ориентированы на более специализированные задачи или структурированы таким образом, чтобы направлять поведение модели в соответствии с ценностями разработчика и ожиданиями пользователей. Модели Frontier AI все чаще становятся мультимодальными: в дополнение к тексту они могут генерировать и обрабатывать другие типы данных (изображения, видео и звук). Ключевыми входными данными для разработки являются вычислительные ресурсы для обучения и запуска модели, данные, на основе которых она может учиться, алгоритмы, определяющие этот процесс обучения, а также таланты и опыт, которые обеспечивают все это [6].

Также отмечено, что ИИ также создает **значительные риски**, что обусловливает необходимость решения вопросов защиты прав человека, прозрачности и объяснимости, справедливости, подотчетности, правового регулирования и безопасности, надлежащего контроля со стороны человека, этики, смягчения предвзятости, конфиденциальности и защиты данных. Отмечены потенциальные непредвиденные риски, связанные со способностью ИИ манипулировать контентом или генерировать вводящий в заблуждение контент. Особые риски безопасности возникают при использовании передового искусственного интеллекта, под которым понимаются те высокомощные модели искусственного интеллекта общего назначения, включая базовые модели, которые могут выполнять широкий спектр задач, а также соответствующие специфические узконаправленные модели ИИ, которые могут демонстрировать возможности, причиняющие вред, которые соответствуют или превосходят возможности, присутствующие в самых передовых моделях сегодняшнего дня. Существенные риски могут возникнуть из-за потенциального преднамеренного неправильного использования или непреднамеренных проблем контроля, связанных с согласованием с намерениями человека. Отчасти эти проблемы связаны с тем, что эти возможности не до конца поняты и поэтому их трудно предсказать. Особую обеспокоенность вызывают риски в сфере кибербезопасности и биотехнологий, а также усиливающиеся риски, связанные с дезинформацией. Существует потенциал для серьезного, даже катастрофического ущерба, преднамеренного или непреднамеренного, вытекающего из наиболее значительных возможностей моделей передового ИИ.

Многие риски, связанные с ИИ по своей природе интернациональный характер, и поэтому необходимо международное сотрудничество, чтобы обеспечить ориентированный на человека, заслуживающий доверия и ответственный искусственный интеллект, который безопасен и служит всеобщему благу. Сотрудничество могло бы включать в себя, где это уместно, классификацию рисков на основе национальных условий и применимых правовых рамок, а также разработку общих принципов и кодексов поведения в области ИИ.

Декларация содержит общее понимание возможностей и рисков, связанных с генеративным искусственным интеллектом (AGI), и понимание настоятельной необходимости осознания потенциальных рисков ИИ и коллективного управления ими посредством новых совместных глобальных усилий по обеспечению безопасной и ответственной разработки и внедрения передового ИИ. Страны-участницы согласились, что существенные риски могут возникнуть в результате потенциального преднамеренного неправильного использования или непреднамеренных проблем с контролем передового ИИ, при этом особую озабоченность вызывают риски кибербезопасности, биотехнологии и дезинформации. Среди основных рисков выделены такие, как предвзятость и нарушение конфиденциальности в применении ИИ. Особое вниманием уделено таким правовым аспектам, как нормативное регулирование передовых технологий, конфиденциальности и защиты данных, а также интеллектуальная собственность.

Другим важным международным событием последнего времени в сфере регулирования искусственного интеллекта стал организованный рядом западных стран так называемый Хиросимский процесс. Группа стран G7 30 октября 2923г. в г. Хиросима (Япония) приняли совместную Декларацию «*G7 Leaders' Statement on the Hiroshima AI Process*» [15], в составе которой приняты два основных документа: свод Международных руководящих принципов по искусственному интеллекту (*The International Guiding Principles on Artificial Intelligence*) и рекомендован Кодекс поведения для разработчиков искусственного интеллекта (*Code of Conduct for AI developers*) [16], содержащий набор правил, которым рекомендуется следовать разработчикам ИИ на добровольной основе для снижения рисков на протяжении всего жизненного цикла ИИ. Хиросимский процесс задуман с целью создания всеобъемлющей политической основы, способствующей разработке безопасных и заслуживающих доверия систем искусственного интеллекта и снижающей риски, возникающие, в частности, от генеративного искусственного интеллекта. Основными пятью рисками признаются: распространение дезинформации и манипулирование, нарушения интеллектуальной собственности, угрозы конфиденциальности, дискриминация и предвзятость, а также риски для безопасности. В Декларации отмечается, что решение задач управления рисками ИИ, исходя из общих принципов верховенства закона и демократических ценностей, требует формирования инклюзивного управления искусственным интеллектом на основе предложенных Международных руководящих принципов и Кодекса поведения для организаций, разрабатывающих передовые системы искусственного интеллекта. Предусматривается, что усилия в сфере ИИ в рамках Хиросимского процесса совместно с Глобальным партнерством по искусственному интеллекту (GPAI) и Организацией экономического сотрудничества и развития (ОЭСР) с участием многих заинтересованных участников, в т.ч. с правительствами, научными кругами, гражданским обществом и частными компаниями не только в странах G7, но и за ее пределами, включая развивающиеся страны и страны с формирующейся рыночной экономикой, будут способствовать созданию открытой и благоприятной среды, в которой безопасные и заслуживающие доверия системы искусственного интеллекта проектируются, разрабатываются, развертываются и используются для максимизации преимуществ технологии при одновременном снижении связанных с ней рисков, для общего блага во всем мире, с целью устранения цифрового разрыва и достижения цифровой инклюзивности.

Предполагается, что свод Международных руководящих принципов по искусственному интеллекту и Кодекс поведения разработчиков ИИ будут постоянно пересматриваться и обновляться, чтобы гарантировать их актуальность, учитывая стремительный характер развития технологий искусственного интеллекта. В Кодексе поведения отмечается, что различные страны могут применять в своей юрисдикции уникальные подходы к реализации правил по-своему. К примеру, для государств Евросоюза такой основой может стать Закон об искусственном интеллекте (*Artificial Intelligence Act*)[17], который, как ожидается, будет принят в начале 2024 года и установит юридически обязательные правила разработки и использования ИИ. Вполне вероятно, что этот Закон создаст образец, по которому страны ЕС будут стремиться моделировать свои собственные законодательные акты в области ИИ.

Принципы служат руководством для организаций, разрабатывающих базовые модели (генеративный) ИИ, и включают следующие 11 ключевых принципов, присущих так называемому «доверенному искусственному интеллекту»:

1. Принятие необходимых мер для выявления, оценки и снижения рисков на протяжении всего жизненного цикла ИИ от разработки до промышленного применения;
2. Выявление и устранение уязвимостей ИИ, включая инциденты и схемы неправильного использования или внедрения, в т.ч. при размещение на рынке;
3. Обнародование сведений о возможностях, ограничениях и областях надлежащего и ненадлежащего использования передовых систем ИИ для поддержания и обеспечения достаточной прозрачности и повышению подотчетности;
4. Применение ответственного обмена информацией и сообщениями об инцидентах среди организаций, разрабатывающих передовые системы ИИ, в т.ч. в промышленности, госуправлении, гражданском обществе и научном сообществе;
5. Разработка, внедрение и раскрытие политики управления ИИ и рисками, основанной на риск-ориентированном подходе, включая политику конфиденциальности и меры по смягчению последствий, в частности для организаций, разрабатывающих передовые системы ИИ;
6. Создание надежных средств контроля безопасности, включая физическую безопасность, кибербезопасность и защиту от внутренних угроз на протяжении всего жизненного цикла ИИ;

15 https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process

16 G7 hiroshima process on generative artificial intelligence (AI): towards A G7 common understanding on generative AI, September 2023, OECD 2023 // http://www.oecd.org/termsandconditions

17 https://artificialintelligenceact.eu/

7. Разработка и внедрение надежных механизмов аутентификации контента и определения происхождения, позволяющие пользователям идентифицировать контент, созданный искусственным интеллектом;
8. Приоритетное внимание исследованиям, направленным на снижение социальных рисков и рисков безопасности ИИ, а также инвестициям в эффективные меры по их снижению;
9. Приоритетное внимание разработке передовых систем искусственного интеллекта для решения глобальных мировых проблем, включая, но не ограничиваясь проблемами климатического кризиса, глобального здравоохранения и образования;
10. Поощрение разработки и принятие международных технических стандартов;
11. Обеспечение мер по вводу данных и защите персональных данных и интеллектуальной собственности.

Единообразие подходов при определении ИИ, общие определения для ИИ на международном уровне и в разных секторах его применения способны обеспечить инклюзивный диалог, устраняя различия между юрисдикциями и способствуя междисциплинарным коммуникациям и сотрудничеству. Основой для таких усилий может служить Концепция Организации экономического сотрудничества и развития (ОЭСР) по классификации систем искусственного интеллекта[18].

ОЭСР обновила свое определение искусственного интеллекта, изложив его в Рекомендациях: «Система искусственного интеллекта – это машинная система, которая для достижения явных или неявных целей на основе получаемых входных данных определяет, как генерировать выходные данные, такие как прогнозы, контент, рекомендации или решения, которые (могут) влиять на физическую или виртуальную среду. Различные системы искусственного интеллекта различаются по уровню автономии и адаптивности после развертывания.»

Концепция определяет, что ***модель искусственного интеллекта*** представляет собой вычислительное представление всей внешней среды системы искусственного интеллекта или ее части, охватывающее, например, процессы, объекты, идеи, людей и/или взаимодействия, которые происходят в этой среде.

Модели искусственного интеллекта используют данные и/или экспертные знания, предоставляемые людьми и/или автоматизированными инструментами, для представления, описания и взаимодействия с реальной или виртуальной средой. При этом, выделяется ИИ «в лабораторных условиях» (*AI «in the lab»*), что относится к концепции и разработке системы ИИ до ее развертывания. Это применимо к данным и входным данным (например, для квалификации данных), модели искусственного интеллекта (например, для обучения исходной модели) и измерениям задачи и выходных данных (например, для задачи персонализации) фреймворка. Это особенно актуально для подходов и требований к управлению рисками *ex ante*. ИИ «в полевых условиях» (*AI «in the field»*) относится к использованию и эволюции системы ИИ после развертывания и применимо ко всем измерениям. Это относится к подходам и требованиям к управлению рисками *ex post*.

Основные характеристики включают технический тип, способ построения модели (с использованием экспертных знаний, машинного обучения или того и другого) и способ использования модели (для каких целей и с использованием каких показателей эффективности). Важно понимание *Жизненного цикла систем искусственного интеллекта*, который включает следующие этапы: (*a*) проектирование, данные и модели, представляющие контекстно-зависимую последовательность, охватывающую планирование и проектирование, сбор и обработку данных, а также построение моделей; (*b*) верификацию и валидацию; (*c*) развертывание; (*d*) эксплуатацию и мониторинг. Эти этапы часто выполняются интерактивным образом и не обязательно являются последовательными. Решение о выводе системы искусственного интеллекта из эксплуатации может быть принято в любой момент на этапе эксплуатации и мониторинга.

В документах OECD[19] сформулированы принципы ответственного управления заслуживающим доверия ИИ, которые дополняют друг друга и должны рассматриваться как единое целое. К ним, в частности, относятся:

- ✓ инклюзивный рост, устойчивое развитие и благосостояние, подразумевающие участие в ответственном управлении заслуживающим доверия искусственного интеллекта в целях расширения человеческих возможностей и креативности, содействия интеграции населения, сокращение экономического, социального, гендерного и других видов неравенства и защита природной среды, тем самым стимулируя инклюзивный рост, устойчивое развитие и благосостояние;

18 OECD framework for the classification of ai systems, OECD 2022. https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm

19 OECD (2023), Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; OECD (2020), Digitalisation and Responsible Business Conduct: Stocktaking of policies and initiatives. https://www.oecd.org/daf/inv/mne/ publicationsdocuments/reports/2/

- ✓ уважение верховенства закона и прав человека (свобода, достоинство и автономия, неприкосновенность частной жизни и защита данных, недискриминация и равенство, многообразие, честность, социальная справедливость и международно признанные трудовые права) и с этой целью внедрение соответствующих механизмов и гарантий, которые соответствуют контексту и соответствуют уровню техники;
- ✓ прозрачность и объяснимость, т.е. ответственное раскрытие значимой информации о системах ИИ, соответствующую контексту и согласующуюся с уровнем техники;
- ✓ надежность и защищенность ИИ на протяжении всего своего жизненного цикла, чтобы в условиях нормального использования, предсказуемого использования или неправильного использования или других неблагоприятных условий они функционировали надлежащим образом и не создавали необоснованного риска для безопасности;
- ✓ подотчетность, т.е. субъекты искусственного интеллекта должны нести ответственность за надлежащее функционирование систем искусственного интеллекта и за соблюдение вышеуказанных принципов, исходя из их ролей, контекста и в соответствии с уровнем техники.

Определение модели ИИ важно для государственной политики, поскольку ключевые свойства моделей ИИ – степень прозрачности и/или объяснимости, надежность и последствия для прав человека, неприкосновенности частной жизни и справедливости – зависит от типа модели, а также от процессов построения модели и логического вывода. Например, системы, использующие нейронные сети, часто рассматриваются как потенциально способные обеспечить сравнительно более высокую точность, но менее объяснимые, чем системы других типов. Объяснимость часто связана со сложностью системы; чем сложнее модель, тем труднее ее объяснить. Степень, в которой модель эволюционирует в ответ на данные, имеет отношение к государственной политике и режимам защиты прав потребителей, особенно для систем искусственного интеллекта, которые могут извлекать уроки из итераций и эволюционировать с течением времени. Понимание того, как была разработана и/или поддерживается модель, является еще одним ключевым фактором при распределении ролей и обязанностей в рамках процессов управления рисками. Субъекты искусственного интеллекта в этом измерении включают разработчиков и моделистов, которые создают, применяют, проверяют и контролируют модели ИИ.

Закон ЕС об искусственном интеллекте классифицирует технологии искусственного интеллекта по трем категориям риска, которые определяют сферу его использования и применения. Во-первых, запрещены к применению технологии и системы ИИ, которые создают неприемлемый риск, такие как государственная система социального скоринга (используемая в КНР). Во-вторых, особые требования установлены законом к технологиям ИИ с высоким уровнем риска, которые могут, к примеру, использоваться как инструмент сканирования резюме в целях ранжирования кандидатов при приеме сотрудников на работу. Остальные модели ИИ, которые явно не запрещены или не перечислены как высокорискованные, в значительной степени остаются нерегулируемыми.

Международный саммит по безопасности ИИ и Хиросимский процесс стали новой вехой в формировании международно-правового регулирования сферы цифровых технологий и прежде всего – искусственного интеллекта как наиболее перспективной и комплексной системы.

Вместе с тем обозначилось отсутствие стремления к установлению консенсуса по вопросам регулирования ИИ на международном уровне, что проявилось в собрании лишь небольшой группы государств (хотя и являющихся лидерами в сфере развития ИИ) для решения исключительно актуального вопроса – создания и развития технологий и моделей генеративного ИИ. Очевидно, решение проблемы безопасности ИИ должно привлекать значительно большее число стран, несомненно заинтересованных в установлении единых правил развития и применения передовых систем искусственного интеллекта.

Выступая на международной конференции «Путешествие в мир искусственного интеллекта AI Journey» 24 ноября 2023 Президент Российской Федерации заявил, что «с внедрением искусственного интеллекта в науку, образование, здравоохранение, все сферы нашей жизни — человечество начинает новую главу своего существования». Указывая на роль искусственного интеллекта сегодня, он отметил, что значение искусственного интеллекта имеет колоссальное значение и о того, каких результатов достигнет страна в соперничестве с другими государствами в сфере искусственного интеллекта, будет зависеть суверенитет, безопасность и состоятельность России. Поставлена задача дальнейшего укоренного развития технологий ИИ для того, чтобы Россия стала одной из самых комфортных юрисдикций для развития искусственного интеллекта. Необходимо разработать на основе генеративного искусственного интеллекта большие отраслевые модели, предложить механизмы их практического внедрения в целях существенного повышения производительности труда, а значит

и заработные платы в ключевых отраслях отечественной экономики. Также предлагается создать глобальные приемлемые для всех правила использования ИИ[20].

Современное понимание ИИ и его безопасность

Проблема безопасности ИИ сформулирована и обсуждается сравнительно недавно, но ее столь широкое правовое оформление на международном уровне сделано впервые. Безопасность искусственного интеллекта представляет собой состояние защищенности от угроз для человека при использовании ИИ, во взаимодействии с ИИ и в системе социальной и биосферы человечества, где ИИ уже стал значимым фактором, оказывающим самостоятельное влияние на общественные отношения, на самого человека. ИИ переходит от стадии инструмента, созданного человеком, к самостоятельной операционной системе, существующей по особым правилам и законам, техническим стандартам, и все чаще на базе машинного обучения (ML) и нейронных сетей.

Как отмечалось в Декларации Хиросимского процесса (2023г.)[21], потенциальные преимущества генеративного ИИ сопряжены с определенными рисками. Способность генеративного ИИ усугублять проблемы дезинформации и манипулирования мнениями рассматривается как одна из основных угроз, исходящих от генеративного ИИ, наряду с рисками нарушения прав интеллектуальной собственности и неприкосновенности частной жизни. Ответственное использование генеративного ИИ, борьба с дезинформацией, защита прав интеллектуальной собственности и управление генеративным ИИ являются одними из главных приоритетов и требуют международного сотрудничества. Другие неотложные и важные вопросы включают конфиденциальность и управление данными, прозрачность, справедливость и предвзятость, права человека и фундаментальные права, безопасность и надежность систем искусственного интеллекта, а также влияние на функционирование демократии.

Экспертами в области ИИ обсуждается вопрос о том, могут ли типы моделей ИИ в конечном итоге привести к созданию искусственному общему интеллекту (AGI), стадии, на которой автономные машины могли бы обладать возможностями человеческого уровня в самых разнообразных вариантах использования. Из-за его потенциального широкого воздействия на общество потенциальные выгоды и риски AGI заслуживают внимания, равно как и потенциально неизбежные последствия узких генеративных систем искусственного интеллекта, которые могут быть столь же значительными, как и AGI.

Долгосрочные преимущества и риски генеративного искусственного интеллекта могут потребовать решений в более широком, системном масштабе, чем уже применяемые подходы к снижению рисков.

Модели генеративного ИИ обучаются на огромных объемах данных, которые включают данные, защищенные авторским правом, в основном без разрешения правообладателей. Продолжающаяся дискуссия о способах защиты интеллектуальной собственности, и в частности, авторских прав, заключается в том, могут ли сами по себе искусственно созданные ИИ результаты интеллектуальной деятельности быть защищены авторским правом или запатентованы, и если да, то кто будет являться правообладателем.

Особый взгляд на ИИ, как отмечено в аналитическом материале OECD[22], связан с распространяемым в цифровых пространствах контентом, где он используется для обучения генеративных моделей искусственного интеллекта, что приводит к порочному негативному циклу в качестве онлайн-информации. Возрастают риски автоматизированных и персонализированных кибератак, слежки и цензуры, чрезмерной зависимости от генерирующих систем, научной чистоплотности разработчиков и концентрации власти и ресурсов.

В долгосрочной перспективе новые формы возможного «асоциального» поведения ИИ предполагают дополнительные риски, включая повышенную активность, стремление к власти и разработку неизвестных подцелей, определяемых машинами для достижения основных целей, запрограммированных человеком, но которые могут не соответствовать человеческим ценностям и намерениям.

Тем не менее, генеративный ИИ быстро внедряется в ключевых секторах промышленности. Прогнозируется, что генеративный ИИ в состоянии создавать существенную экономическую ценность и социальное благополучие. Компании начали внедрять технологии для создания новых бизнес-возможностей, а стартапы конкурируют за венчурный капитал. Популярные на сегодняшний день варианты использования программ-приложений включают предварительную обработку данных, сжатие и классификацию изображений, медицинскую визуализацию, персонализацию и интуитивно понятные пользовательские интерфейсы.

Системы генеративного искусственного интеллекта (ИИ) создают новый контент в ответ на запросы, основанные на их обучающих данных. Распространение систем генеративного искусственного интеллекта

20 http://www.kremlin.ru/events/president/transcripts/72811

21 G7 hiroshima process on generative artificial intelligence (ai): towards a G7 common understanding on generative AI, September 2023, OECD 2023. http://www.oecd.org/termsandconditions

22 Initial policy considerations for generative artificial intelligence, oecd artificial intelligence papers, September,2023. URL: http://www.oecd.org/

высветили возможности искусственного интеллекта, включая, например, ChatGPT для текста или для изображений (Stable Diffusion), для аудио или видео (DeepVoice), а также мультимодельные системы, объединяющие несколько типов медиа или языковые модели ИИ.

Однако эти же технологии создают критические социальные и политические проблемы, которые выражаются в потенциальных изменениях на рынках труда, неопределенности в отношении прав интеллектуальной собственности; риски, связанные с возможностью злоупотреблений при создании дезинформации и манипулируемого контента, распространением ложной информации (*deep fake*). В итоге могут формироваться негативные социальные, политические и экономические последствия, включая дезинформацию по ключевым научным вопросам, создание стереотипов и дискриминации, искажение общественного дискурса, создание и распространение теорий заговора и другой дезинформации, влияние на политические выборы, искажение рыночной информации и даже подстрекательство к насилию. Это, тем не менее, не отрицает преобразующую природу генеративного искусственного интеллекта и значимости международных дискуссий о стремлении к инклюзивному и заслуживающему доверия искусственному интеллекту.

Вместе с тем генеративный искусственный интеллект значительно увеличивает масштабы и объем дезинформации. В 2022 году было обнаружено, что люди почти в 50% случаев неспособны отличить искусственный интеллект от новостей, созданных человеком. Это означает, что генеративный ИИ может усилить риски как дезинформации (непреднамеренного распространения ложной информации), так и преднамеренной дезинформации злоумышленниками. Передовые модели генеративного искусственного интеллекта обладают мультимодальными возможностями, которые могут усугубить эти риски, например, путем объединения текста с изображением, видео или даже голосами[23].

Еще одной тревожной особенностью технологий ИИ является их склонность к «галлюцинациям» (т.е. к получению неверных, но убедительных результатов), особенно когда ответ отсутствует в данных обучения. Это позволяет может создавать убедительную дезинформацию, разжигать ненависть или воспроизводить предубеждения. Риски также включают чрезмерное доверие и чрезмерную зависимость от модели ИИ, что приводит к зависимости, которая может помешать развитию навыков и даже привести к потере навыков (OpenAI).

Синтетический контент может быть особенно полезен в политике, науке и правоохранительных органах. Риски, связанные с моделями искусственного интеллекта, генерирующими текст в изображение, ясно показывают, насколько быстр технологический прогресс.

Многочисленные «фотографии» в Twitter и других онлайн-платформах изображали известных политических деятелей и глав государств, совершающих неожиданные действия, но при этом вызывали большое доверие, демонстрируя силу синтетических образов, особенно в поляризованных политических контекстах. Другой проблемой было манипулирование научными изображениями для создания ложной информации (*deep fake*), угрожающей доверию в исследовательских и научных сообществах, а также репутации науки в глазах широкой общественности. В качестве примеров можно привести использование синтетических изображений отрицателями изменения климата или распространение дезинформации о COVID-19[24].

Вместе с тем выделяются и наиболее важные приоритеты ИИ, включая права человека и основные свободы, безопасность и надежность систем искусственного интеллекта, демократические ценности, а также конфиденциальность и управление данными (см. Табл. 1)

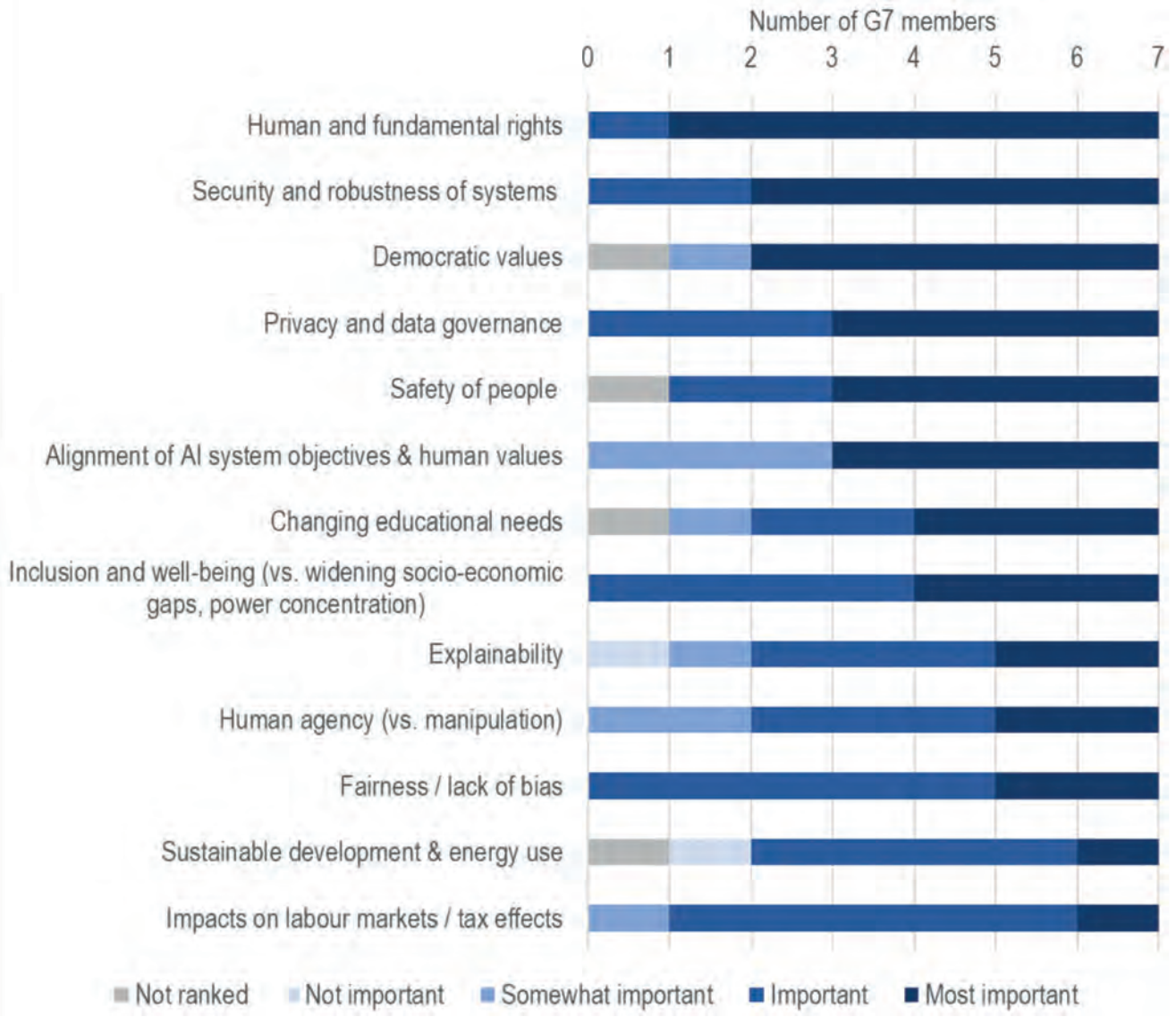
В последние годы все большее применение технологии искусственного интеллекта используют в целях упрощения, автоматизации и ускорения работы в различных областях [8], включая услуги по автоматизации заключения и мониторинга контрактов, программное обеспечение для электронного раскрытия информации, программные продукты для управления делами, агрегаторы информации, позволяющие принимать более обоснованные решения при выборе бизнес-решений или решений для прогнозирования судебных процессов. При этом, все острее встают вопросы безопасности и правомерности действий с учетом проблем имитации и ложных фактов, фальсифицированных и фальшивых фактов (Deep fake), создаваемых с помощью тех же технологий искусственного интеллекта. Предполагается, что решением этих проблем могло бы стать создание цифровой метки происхождения, т.е. встраивание цифровых «отпечатков пальцев» в соответствующие носители, с использованием возможностей технологии блокчейн. Мир искусства уже принял аналогичные процедуры, изобретательно и очень прибыльно используя так называемые NFT (*non-fungible tokens – невзаимозаменяемые токены*) для подтверждения

23 Initial policy considerations for generative artificial intelligence, oecd artificial intelligence papers, September,2023. URL: http://www.oecd.org/

24 Initial policy considerations for generative artificial intelligence, oecd artificial intelligence papers, September, 2023. URL: http://www.oecd.org/

Таблица 1

G7 Hiroshima process on artificial intelligence (AI)



уникальности произведений искусства, хранящихся в цифровом виде.

Как отмечают эксперты [9,10], в условиях обеспечения технологического суверенитета страны при усилении наступающей составляющей информационной безопасности идет совершенствование таксономий, развитие технологий и средств защиты, их адаптация под новые архитектуры (облачные, микросервисные и пр.), внедрение прорывных технологий (AI/ML, Big Data), а также формирование новых требований применения программного обеспечения с открытым исходным кодом, особенно в условиях новых угроз и рисков.

В этой связи целесообразно предусмотреть и согласовать принципы правового регулирования и юридически значимые этические принципы использования технологий искусственного интеллекта, внедрить стандартизированные правила его использования в правосудии и разрешении коммерческих споров, включая развитие современной цифровой криминалистики. Искусственный интеллект важен, в т.ч., с точки зрения понимания его функциональности и безопасности, а также соответствия принципу верховенства права и подконтрольности решений ИИ человеку [10].

Передовые высокоэффективные базовые модели ИИ могут обладать опасными возможностями, достаточными для создания серьезных рисков для общественной безопасности и регулирования. Для решения этих проблем необходимы по крайней мере три составных элемента регулирования моделей «*Frontier AI*»: (a) установление стандартов и соответствующих требований к разработчикам моделей ИИ; (b) установление требований и регистрации и отчетности для обеспечения регулирующим органам наглядность процессов разработки модели ИИ,

и (с) механизмы для обеспечения соответствия стандартам безопасности при разработке и внедрении моделей *Frontier AI*. Саморегулирование отрасли является важным первым шагом. Однако для создания стандартов безопасности и обеспечения их соответствия необходимо предоставить надзорным органам полномочий по обеспечению контроля соблюдения стандартов и режим лицензирования моделей *Frontier AI*.

Первоначальный набор стандартов безопасности ИИ включает: проведение оценок рисков перед развертыванием; внешний контроль поведения модели; использование оценок рисков для обоснования решений о развертывании, а также мониторинг и реагирование на новую информацию о возможностях модели и ее использовании после развертывания. *Frontier AI* способен выполнять широкий спектр задач и дополняется инструментами для расширения своих возможностей. Прогресс в течение следующих нескольких лет может быть быстрым и неожиданным в определенных отношениях. Вполне возможно, что в недалеком будущем могут быть разработаны продвинутые агенты ИИ общего назначения, но существует несколько глубоких, нерешенных сквозных технических и социальных факторов риска развития ИИ.[25]

Вопросы обеспечения международной информационной и кибербезопасности безопасности стали особенно актуальны как в практическом плане, так и в науке международного права в условиях нарастающих вызовов и угроз, связанных с использованием современных технологий в т. ч. против суверенитета государств, осуществления в глобальном информационном пространстве действий, препятствующих поддержанию международной безопасности и стабильности.

В связи с этим, в соответствии с Концепцией внешней политики Российской Федерации приоритетами России в целях обеспечения международной информационной безопасности, противодействия угрозам в ее отношении, укрепления российского суверенитета в глобальном информационном являются:

- ✓ укрепление и совершенствование международно-правового режима предотвращения и разрешения межгосударственных конфликтов и регулирования деятельности в глобальном информационном пространстве;
- ✓ формирование и совершенствование международно-правовых основ противодействия использованию информационно-коммуникационных технологий в преступных целях;
- ✓ обеспечение безопасного и стабильного функционирования и развития информационно-телекоммуникационной сети «Интернет» на основе равноправного участия государств в управлении данной сетью и недопущению установления иностранного контроля над ее национальными сегментами;
- ✓ принятие политико-дипломатических и иных мер, направленных на противодействие политике недружественных государств по милитаризации глобального информационного пространства, по использованию информационно-коммуникационных технологий для вмешательства во внутренние дела государств и в военных целях, а также по ограничению доступа других государств к передовым информационно-коммуникационным технологиям и усилению их технологической зависимости.

Такой подход отличается от стратегий западных стран, ограничивающихся рамками «кибербезопасности», представляющей собой свод процессов, передовых практик и технологий, которые помогают защитить критически важные системы и сети от цифровых атак.

Россия выступает за широкий подход к содержанию данного понятия, включая в него как технические аспекты (безопасность цифровых технологий, информационных систем и сетей), так и значительный круг политико-идеологических вопросов (манипулирование информацией, пропаганда в глобальных информационных сетях, информационное воздействие). Страны коллективного Запада и США придерживаются узкого подхода, ограничиваясь технической стороной, используя термин «кибербезопасность». Учитывая специфику современных информационных отношений, многоаспектность рисков и угроз в информационном пространстве, позицию России следует признать более обоснованной и нацеленной на системный подход в вопросах регулирования обеспечения информационной безопасности.

В условиях цифровизации возникают новые информационные вызовы и угрозы, многие из которых носят трансграничный (глобальный) характер [11]. Как отмечается в докладе экспертов МГИМО, предложенный и продвигаемый Россией термин «международная информационная безопасность» подразумевает наличие не только технических, но и политико-идеологических угроз в данной области, что не коррелирует с западной концепцией кибербезопасности, акцентирующей внимание на технологическом измерении информационных угроз [12].

В России принято несколько правовых документов в сфере регулирования искусственного интеллекта,

25 Capabilities and risks from frontier AI, AI Safety Summit, October 2023. URL: https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/ frontier-ai-capabilities-risks-report.pdf.

в которых вопросам безопасности ИИ отводится важное место. Так, в 2019 году утверждена специальная Национальная стратегия развития искусственного интеллекта на период до 2030 года[26], которая провозглашает целями развития искусственного интеллекта в Российской Федерации обеспечение роста благосостояния и качества жизни ее населения, обеспечение национальной безопасности и правопорядка, достижение устойчивой конкурентоспособности российской экономики, в том числе лидирующих позиций в мире в области искусственного интеллекта.

Литература

1. *Ghazinoory, S., Fatemi, M., Saghafi, F. et al. A Framework for Future-Oriented Assessment of Converging Technologies at National Level. Nanoethics 17, 8 (2023). https://doi.org/10.1007/s11569-023-00435-4.*
2. *Мохов А. А. Демографическая безопасность и ее правовое обеспечение // Юрист. 2023. N6. С. 62–67.*
3. *Amatova, N. E.: Social consequences of the implementation of NBIC-technologies: risks and expectations. Univ. Soc. Sci. 9(8) (2014). http://7universum.com/ en/social/ archive/ item/1549. Accessed 22 Jan 2020.*
4. *S. Klaus, C. Jung. Legal Aspects of «Artificial Intelligence» (AI) / Information and Communication Technology Newsletter, 2019, N10. https://www.swlegal.com/media/filer_public/ce/e4/cee498cc-910d-4af8-a020-5b4063662b35/sw_newsletter_october_i_english.pdf*
5. *Haskins A., Arora S., Nilawar U. Impact of Artificial Intelligence on Indian Real Estate: Transformation Ahead // Colliers radar Property Research (India). 05.10.2017. 13 p. P. 4.;*
6. *Capabilities and risks from frontier AI, AI Safety Summit, 2023. URL: https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf;*
7. *Frontier AI Regulation: Managing Emerging Risks to Public Safety, November 7, 2023.URL: https://arxiv.org/abs/2307.03718*
8. *The Paradox of Artificial Intelligence in the Legal Industry: Both Treasure Trove and Trojan Horse? // The Perils of Deepfakes, Wolters Kluwer. 2021 // URL: http://arbitrationblog.kluwerarbitration.com.*
9. *Марков А. С. Важная веха в безопасности открытого программного обеспечения// Вопросы кибербезопасности, 2023, №1(53), С.2–12*
10. *Карцхия А. А. LegalTech как основа цифровой правовой экосистемы / LegalTech в сфере предпринимательской деятельности: монография (отв. ред. И.В. Ершова, О.В. Сушкова), М: Проспект, 2023. С.25–33*
11. *Карцхия А. А., Макаренко Г. И., Макаренко Д. Г. Правовые перспективы технологий искусственного интелекта // Безопасные информационные технологии / Сборник трудов Двенадцатой международной научно-технической конференции МВТУ им Н. Э. Баумана. 2023. С. 154–161.*
12. *Крутских А. В., Зиновьева Е. С. Международная информационная безопасность: подходы России. М.: МГИМО МИД России, 2021. С. 6.*

26 Утв. Указ Президента РФ от 10.10.2019 N 490 «О развитии искусственного интеллекта в Российской Федерации» // Собрание законодательства РФ, 14.10.2019, N 41, ст. 5700

СПЕЦИАЛЬНАЯ МОДЕЛЬ БЕЗОПАСНОСТИ СОЗДАНИЯ И ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Гарбук С. В.[1]

DOI: 10.21681/2311-3456-2024-1-15-23

Цель: *обоснование характера и структуры угроз, проявляющихся при создании и применении систем искусственного интеллекта на базе алгоритмов машинного обучения, в зависимости от видов нарушений требований, предъявляемых к информационным компонентам этих систем.*

Методы: *исследования проводились с использованием методов формальной логики, многокритериальной оценки, инженерии программных систем.*

Результат: *показано, что угрозы безопасности создания и применения систем искусственного интеллекта обусловлены нарушением требований в области целостности и доступности, предъявляемых к функциональным характеристикам систем и предусмотренным условиям их эксплуатации, к используемым эталонным архитектурам моделей машинного обучения, а также к обучающим, тестовым и входным наборам данных. При этом угрозы безопасности могут проявляться в виде деградации и повышения погрешности оценки функциональных характеристик (нарушение функциональности), а также компрометации чувствительных сведений о самих системах и о третьих лицах. Нарушение функциональности систем, в свою очередь, может приводить к реализации угроз физической, информационной и экономической безопасности. В статье приведены логические зависимости, позволяющие оценивать структуру потенциальных угроз безопасности в зависимости от степени соответствия комплексу предъявляемых требований в области целостности и конфиденциальности информационных компонент систем.*

Научная новизна: *полученные результаты могут быть использованы при обосновании требований к процессам жизненного цикла систем искусственного интеллекта на основе алгоритмов машинного обучения, а также при оценивании потенциальных рисков при создании и применении таких систем.*

Ключевые слова: *качество систем искусственного интеллекта, функциональная корректность систем искусственного интеллекта, риски создания и применения систем искусственного интеллекта.*

A SPECIAL SECURITY MODEL FOR THE CREATION AND APPLICATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

Garbuk S. V.[2]

The goal of the investigation: *to substantiate the nature and structure of threats manifested in the creation and application of artificial intelligence systems based on machine learning algorithms, from the types of violations of the requirements imposed on the information components of these systems.*

Methods: *The research was conducted using methods of formal logic, multi-criteria evaluation, and software systems engineering.*

Result: *It is shown that the threats to the security of the creation and application of artificial intelligence systems are caused by violation of the requirements in the field of integrity and accessibility imposed on the functional characteristics of systems and the provided conditions of their operation, to the used reference architectures of machine learning models, as well as to training and test datasets. At the same time, security threats can manifest themselves in the form of degradation and increased error in evaluating functional characteristics (violation of functionality), as well as compromising sensitive information about the systems themselves and about third parties. Violation of the functionality of systems, in turn, can lead to the realization of threats to physical, information and economic security. The article presents logical dependencies that allow us to assess the composition of potential security threats depending on the degree of compliance with the set of requirements in the field of integrity and confidentiality of information components of systems.*

1 Гарбук Сергей Владимирович, кандидат технических наук, старший научный сотрудник, НИУ ВШЭ, г. Москва, Россия. ORCHID: 0000-0001-5385-3961, E mail: sgarbuk@hse.ru .

2 Sergey V. Garbuk, Ph.D., Senior Research Fellow, National Research University Higher School of Economics, Moscow, Russia. ORCHID: 0000-0001-5385-3961, E mail: sgarbuk@hse.ru.

Scientific novelty: *The results obtained can be used to substantiate the requirements for the life cycle processes of artificial intelligence systems based on machine learning algorithms, as well as to assess potential risks in the creation and application of such systems.*

Keywords: *Security of artificial intelligence systems, quality of artificial intelligence systems, functional correctness of artificial intelligence systems, risks of creation and application of artificial intelligence systems.*

Введение

Впечатляющие результаты в области технологий искусственного интеллекта (ИИ) на современном этапе обусловлены, прежде всего, развитием одного из направлений ИИ – алгоритмов машинного обучения (МО), обеспечивающих эффективное решение различных задач в области автоматизированной обработки данных в условиях отсутствия основанных на знаниях моделей наблюдаемых объектов и процессов. Наряду с универсальностью применения алгоритмы МО обладают также такими особенностями, как отсутствие полной интерпретируемости, возможность дообучения алгоритмов МО в процессе эксплуатации систем ИИ (СИИ), высокая актуальность вопросов социальной приемлемости применения СИИ, необходимость сравнения функциональных возможностей СИИ и человека-оператора и др. [1].

Модель жизненного цикла и требования к информационным компонентам систем искусственного интеллекта

Модель жизненного цикла (ЖЦ) для типовой системы ИИ, учитывающая приведенные выше особенности алгоритмов МО, представлена на рис. 1 [2]. На разных этапах ЖЦ СИИ используются различные информационные компоненты (данные, формализованные описания, модели), необходимые для успешной реализации этих этапов:

- ✓ функциональные требования к системам (ФТ), которые могут быть представлены в виде вектора $F = \{F_1, F_2, \ldots F_N\}$ [3];
- ✓ описание предусмотренных условий эксплуатации (ПУЭ), в общем случае заданное многомерной плотностью распределения $W(E)$ существенных факторов эксплуатации (СФЭ) СИИ $E = \{e_1, e_2 \ldots e_K\}$ [3], где K – количество СФЭ;
- ✓ эталонные архитектуры программного обеспечения, реализующего алгоритмы МО (требования к этой информационной компоненте могут быть представлены в виде набора R_R);
- ✓ обучающие и дообучающие НД (R_{L1} *и* R_{L2}, соответственно);
- ✓ тестовые НД ($R_{T1} = R_{T2} = R_T$);
- ✓ входные данные (R_U).

Отметим, что подобный набор информационных компонент специфичен для СИИ, так что особенности систем обработки данных на основе алгоритмов МО полностью обуславливаются этим набором как минимально достаточным. При этом в составе СИИ могут присутствовать иные компоненты (например сенсоры, средства передачи, обработки, хранения и отображения информации, исполнительные устройства и т.п), также определяющие качество и безопасность работы системы, но не являющихся специфическими для систем обработки данных на основе алгоритмов МО.

Таким образом, полный набор требований к СИИ может быть представлен в виде объединения: $R_C \cup F \cup W(E) \cup R_R \cup R_{L1} \cup R_T \cup R_U \cup R_{L2}$, где R_C – набор общих требований к СИИ, не зависящих от особенностей используемых алгоритмов МО; R – множество требований, специфичных для СИИ, которое для удобства можно переписать в виде: $R = R_1 \cup R_2 \cup R_3 \cup R_4 \cup R_5 \cup R_6 \cup R_7$ (индексы требований соответствуют обозначениям информационных компонент на рис. 1).

Тогда, если система S обладает множеством специфичных для алгоритмов МО характеристик $r(S) = \{r_n(S)\}_{n=1..7}$, то под специальной моделью безопасности СИИ будем понимать совокупность зависимостей, определяющих влияние несоответствий характеристик r системы S требованиям из набора R на характер угроз, проявляющихся при создании и применении этой системы.

Будем считать также, что требования к информационным компонентам из набора R могут быть двух видов:

- ✓ требования целостности (используется индекс in, где n – номер информационной компоненты, $n = 1..7$, рис. 1), заключающиеся в корректности формирования информационной компоненты и в предотвращении её умышленных и непреднамеренных искажений;
- ✓ требования конфиденциальности (cn), заключающиеся в предотвращении компрометации соответствующих данных.

Далее будут рассмотрены наиболее типичные примеры, иллюстрирующие негативные последствия, вызванные отклонением от выполнения тех или иных требований. При этом будет приниматься во внимание, что среди характеристик СИИ $r(S)$ следует выделять как целевые характеристики (прежде всего, это ФХ СИИ $r_1(S) = f(S)$, подтвержденные в конкретных условиях эксплуатации $r_2 = w(E)$, а также характеристики конфиденциальности самой СИИ и прочих субъектов информационной безопасности),

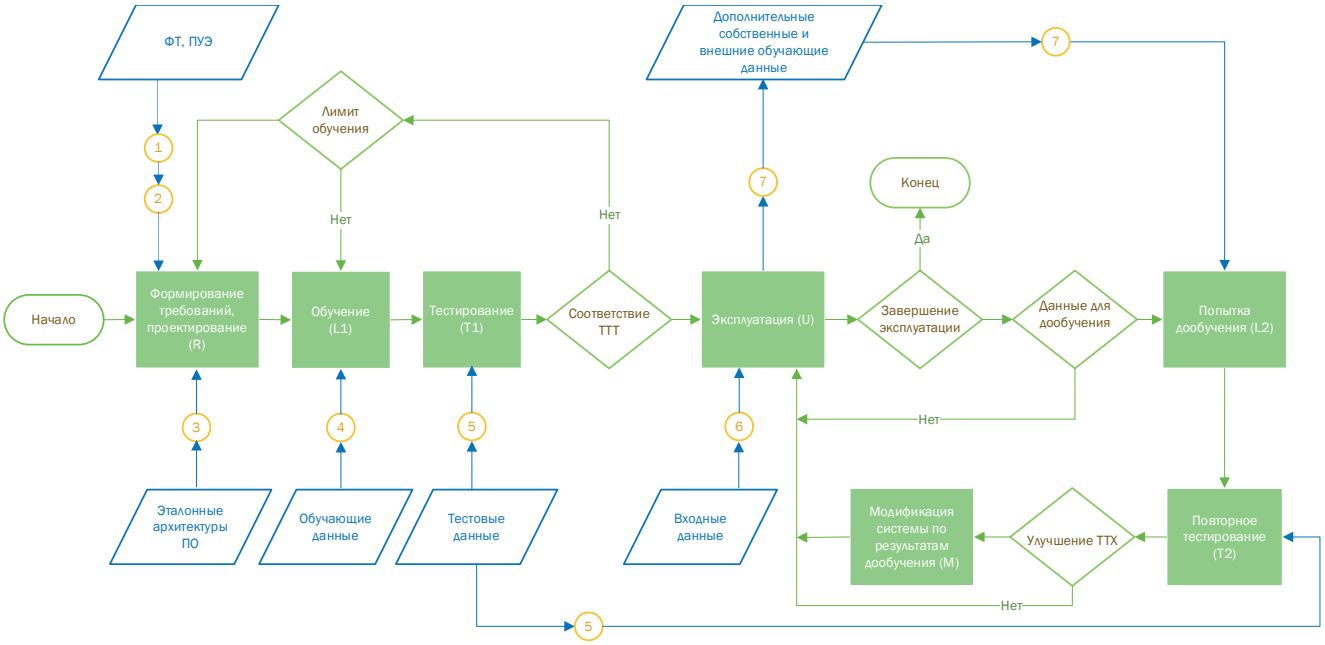


Рис.1. Модель ЖЦ СИИ. Цифрами на рисунке обозначены точки возникновения несоответствия требованиям, установленным к различным информационным компонентам

так и «процессные» характеристики, важные не сами по себе, а лишь в силу влияния этих характеристик на целевые (например, характеристики целостности и конфиденциальности используемых НД).

Угрозы безопасности при создании и применении систем искусственного интеллекта

Нарушение целостности ФТ (нарушение требований вида $i1$, то есть $r_{i1} \neq R_{i1}$) может быть обусловлено недостаточной представительностью выбранного набора существенных характеристик и показателей качества СИИ[3], некорректными весовыми коэффициентами частных показателей качества, влияющих на точность определения интегрального показателя, неправильно выбранными пороговыми значениями показателей качества. В свою очередь, нарушение конфиденциальности ФТ (нарушение требований $c1$) является нежелательным при наличии активного злоумышленника.

Выполнение требований целостности к предусмотренным условиям эксплуатации (ПУЭ) СИИ (требования вида $i2$) обеспечивает возможность формирования репрезентативных обучающих НД и выбора оптимальных архитектур СИИ, проведения репрезентативных испытаний системы, а также возможность принятия обоснованных решений по применению СИИ. Таким образом, нарушение выполнения требований $i2$ будет приводить как к непосредственной деградации функциональных характеристик (ФХ) СИИ, так и к нарушению целостности выбранных архитектур систем и обучающих НД, что может быть записано в виде:

$$(r_{i2} \neq R_{i2}) \rightarrow (r_{i3} \neq R_{i3}); (r_{i2} \neq R_{i2}) \rightarrow (r_{i4} \neq R_{i4}). \quad (1)$$

Наиболее опасным является нарушение требований к целостности входных данных, обрабатываемых в СИИ (требования $i6$). Создаваемые системы ИИ, в частности – реализованные на искусственных нейронных сетях (ИНС), оказываются неустойчивы к небольшим искажениям входных данных. Недостаточное выполнение требований i6 приводит к завышению оценок ФХ СИИ, так как в процессе испытания состязательные атаки как фактор, снижающий качество работы системы, учитывается недостаточно.

Невыполнение требований к конфиденциальности сведений об использованных архитектурах СИИ (несоответствие требованиям $c3$), обучающих ($c4$), тестовых ($c5$) и дообучающих ($c6$) НД создает дополнительные риски невыполнения требования к целостности входных данных [4, 5]:

$$(r_{c3} \neq R_{c3}) + (r_{c4} \neq R_{c4}) + (r_{c5} \neq R_{c5}) + (r_{c7} \neq R_{c7}) \rightarrow \rightarrow (r_{i6} \neq R_{i6}). \quad (2)$$

В процессе работы СИИ уровень конфиденциальности данных, накапливаемых и обрабатываемых в системе, может возрастать, что потенциально может привести к неправильному определению (занижению) требований в области ИБ на стадии проектирования СИИ. Кроме того, учитывая, что входные данные с высоким уровнем конфиденциальности используются в данном случае и для дообучения

3 ГОСТ Р 59898–2021 Оценка качества систем искусственного интеллекта. Общие положения. Введен 2022 03-01. М.: Российский институт стандартизации. 2021, 20 с.

СИИ, нарушение требований вида $c1$ создаст предпосылки для успешной реализации атаки на конфиденциальность обучающих данных СИИ:

$$(r_{c1} \neq R_{c1}) \rightarrow (r_{c4} \neq R_{c4}), (r_{c1} \neq R_{c1}) \rightarrow (r_{c7} \neq R_{c7}). \quad (3)$$

Нарушение требований к целостности архитектур СИИ и обучающих НД в дальнейшем повышает эффективность реализации атак на целостность входных данных:

$$(r_{i3} \neq R_{i3}) + (r_{i4} \neq R_{i4}) \rightarrow (r_{i6} \neq R_{i6}). \quad (4)$$

Предотвращение атак на целостность архитектуры СИИ и обучающих НД заключается, прежде всего, в использовании эталонных моделей МО из доверенных источников [6], а также в реализации разработчиком и эксплуатантом (в случае дообучения СИИ в процессе эксплуатации) необходимого комплекса мер информационной безопасности.

Невыполнение требований к целостности тестовых, как и обучающих, НД (требования вида $i5$ может приводить к нарушению их репрезентативности, точности, достоверности и к реализации других факторов снижения качества. При этом функциональные характеристики систем ИИ не изменяются, но оценка этих характеристик, получаемая при испытании систем с использованием модифицированных тестовых НД, может содержать недопустимо высокую погрешность. Следует различать две составляющие погрешности оценки ФХ СИИ:

- ✓ смещение характеристик, вызванное нарушением баланса тестовых примеров различной сложности. Относительная нехватка сложных примеров приводит к неоправданному завышению (чрезмерно оптимистичной оценке) ФХ, избыток сложных примеров – к недооценке ФХ;
- ✓ возрастание случайной составляющей погрешности, обусловленной недостаточным объёмом тестового НД.

Отметим, что возникновение предпосылок к завышению функциональных возможностей испытуемых алгоритмов МО, может объясняться как естественными причинами (например, недостаточной квалификацией персонала, подготавливающего тестовые НД), так и умышленными действиями злоумышленника. В обоих случаях наличие существенных погрешностей в оценке функциональных характеристик может привести к некорректному применению СИИ по назначению, а в случае умышленных искажений – еще и создать дополнительные предпосылки для реализации противником эффективных атак на входные данные:

$$(r_{i5} \neq R_{i5}) \rightarrow (r_{i6} \neq R_{i6}). \quad (5)$$

Таким образом, анализ рисков, проявляющихся при невыполнении требований типа in и cn, показал, что возможными негативными последствиями, обусловленными несоответствием требованиям, специфичных для СИИ на основе алгоритмов МО, являются:

1) существенное возрастание ошибки оценивания ФХ при тестировании (испытаниях) СИИ за счет смещения (как правило – в сторону завышения характеристик) и возрастания случайной составляющей погрешности оценок при снижении вариативности тестовых НД. Неточное понимание функциональности систем существенно усложняет или даже делает невозможным принятие эксплуатирующей стороной рациональных решений об использовании СИИ на практике;
2) деградация ФХ, ограничивающая возможность применения систем в реальных условиях эксплуатации. Причины такой деградации заключаются либо во внесении преднамеренных искажений в обучающие НД и архитектуру СИИ, в результате чего ФХ ухудшаются в предусмотренных условиях эксплуатации, либо в создании злоумышленниками в ходе реального применения СИИ условий применения, существенно отличающихся от предусмотренных разработчиками системы (ПУЭ). Во втором случае ФХ систем сохраняют гарантированные разработчиком значения в предусмотренных условиях эксплуатации, однако деградируют в реальных условиях, выходящих за рамки ПУЭ;
3) нежелательное нарушение конфиденциальности сведений о тактико-технических характеристиках и особенностях применения СИИ, приводящее, например, к повышению эффективности деструктивных информационных воздействий на СИИ злоумышленниками, в том числе – за счет оптимизации способов искажения входных данных СИИ и т.п.;
4) компрометация сведений о физических и юридических лицах, интересы которых так или иначе затрагиваются при реализации процессов ЖЦ СИИ (заинтересованные лица СИИ).

Зависимости возможных негативных последствий, обусловленных нарушением специальных требований при создании и применении СИИ, представлены в табл. 1. В таблице использованы обозначения inm и cnm для зависимостей, характеризующих влияние нарушения соответственно целостности и конфиденциальности n-й информационной компоненты на m-е негативное последствие, $m = 1..4$.

Таблица 1.

Матрица специальных требований к информационным компонентам СИИ и возможных негативных последствий, обусловленных невыполнением этих требований

Объект управления качеством и безопасностью (информационная компонента СИИ)	Аспект управления качеством и безопасностью (вид требований к информационной компоненте)	Возможные негативные последствия от несоответствия требованиям			
		1. Рост погрешности оценки ФХ при тестировании СИИ	2. Деградация ФХ СИИ в ПУЭ	3. Компрометация ТТХ и вариантов применения СИИ	4. Компрометация данных о заинтересованных лицах СИИ
1. Функциональные требования (ФТ) к СИИ	$i1$ – полнота набора и обоснованность критериальных значений (в том числе – путем оценки возможностей человека-оператора) ФТ, обоснованность метрик для сравнения функциональных характеристик (ФХ) с ФТ	$i11$ – смещение оценки интегрального показателя качества	$i12 = 0$	$i13 = 0$	$i14 = 0$
	$c1$ – конфиденциальность ФТ	$c11 = 0$	$c12 = 0$	$c13$ – прямая компрометация ТТХ	$c14 = 0$
2. Описание предусмотренных условий эксплуатации (ПУЭ)	$i2$ – полнота набора существенных факторов эксплуатации (СФЭ), соответствие законов распределения СФЭ в ПУЭ и в реальных условиях применения СИИ	$i21$ – смещение и возрастание случайной составляющей погрешности оценки ФХ[41]	$i22$ – ухудшение ФХ вследствие неверного выбора обучающих НД и архитектур СИИ	$i23 = 0$	$i24 = 0$
	$c2$ – конфиденциальность ПУЭ, предотвращение компрометации сведений, которые могут быть использованы злоумышленником для целенаправленного создания условий, выходящих за пределы ПУЭ	$c21$ – смещение оценки ФХ	$c22$ – ухудшение ФХ при создания злоумышленником условий, выходящих за ПУЭ	$c23$ – компрометация ТТХ в части ПУЭ	$c24 = 0$
3. Архитектура СИИ	$i3$ – отсутствие программных закладок, обеспечивающих злоумышленнику повышенные по сравнению с предполагавшимися при тестировании возможности по реализации информационных атак на входные данные (например, по подбору эффективных состязательных атак)	$i31$ – смещение оценки ФХ	$i32$ – ухудшение ФХ	$i33 = 0$	$i34 = 0$
	$c3$ – конфиденциальность сведений об архитектуре СИИ, которые могут быть использованы для реализации эффективных атак на входные данные	$c31$ – смещение оценки ФХ	$c32 = 0$	$c33$ – раскрытие уязвимостей к атакам на входные данные	$c34 = 0$

4 Смещение оценки функциональных характеристик, вызванное несоответствием установленным к СИИ требованиям, как правило приводит к завышению (получению излишне оптимистичных) оценок.

4. Обучающие НД	$i4$ – предотвращение целенаправленных искажений («отравления») обучающих НД, приводящих к повышенным по сравнению с предполагавшимися при тестировании возможностям злоумышленника по реализации информационных атак на входные данные (например, по подбору эффективных состязательных атак)	$i41$ – смещение оценки ФХ	$i42$ – ухудшение ФХ	$i43 = 0$	$i44 = 0$
	$c4$ – конфиденциальность обучающих НД, предотвращение компрометации сведений, облегчающих последующую реализацию эффективных атак на входные данные	$c41$ – смещение оценки ФХ	$c42 = 0$	$c43$ – раскрытие ПУЭ (вариативность обучающего НД) и уязвимостей к атакам на входные данные	$c44$ – раскрытие конфиденциальных данных по объектам, на которых происходит обучение
5. Тестовые НД	$i5$ – предотвращение искажений, влияющих на репрезентативность тестового НД – смещение (нарушение баланса тестовых сценариев разной сложности) и снижение вариативности (сокращение объема выборки, большое число дублирующих тестовых сценариев)	$i51$ – смещение и возрастание случайной составляющей погрешности оценки ФХ	$i52 = 0$	$i53 = 0$	$i54 = 0$
	$c5$ – предотвращение компрометации тестовых НД, способной привести к снижению представительности тестирования (доступ к тестовым НД разработчиков приводит к переобучению СИИ, доступ злоумышленников – повышению возможностей по реализации информационных атак на входные НД	$c51$ – смещение оценки ФХ	$c52 = 0$	$c53$ – раскрытие ПУЭ (вариативность тестового НД) и уязвимостей к атакам на входные данные	$c54$ – раскрытие конфиденциальных данных по объектам, на которых сформирован тестовый НД
6. Входные данные СИИ	$i6$ – защита от искажений, в том числе – умышленных (например, состязательные атаки) и естественных	$i61$ – смещение оценки ФХ	$i62 = 0$	$i63 = 0$	$i64 = 0$
	$c6$ – конфиденциальность входных данных, в том числе, с учетом возрастания уровня их конфиденциальности при накоплении	$c61 = 0$	$c62 = 0$	$c63$ – раскрытие сценариев применения СИИ	$c64 = 0$

7. Дообучающие НД	$i7$ – достоверность и информативность дообучающих НД, предотвращение статистического смещения дообучающего НД, вызванного существенным отклонением условий эксплуатации СИИ от предусмотренных	$i71$ – смещение оценки ФХ для ПУЭ	$i72$ – ухудшение ФХ в ПУЭ	$i73 = 0$	$i75 = 0$
	$c7$ – конфиденциальность дообучающих НД, предотвращение компрометации сведений, облегчающих последующую реализацию эффективных атак на входные данные	$c71$ – смещение оценки ФХ	$c72 = 0$	$c73$ – раскрытие реализованных сценариев применения СИИ	$c74$ – раскрытие данных по объектам дообучения СИИ

Проиллюстрированные в табл. 1 зависимости характера потенциального ущерба от вида нереализуемых требований к информационным компонентам СИИ могут быть использованы для прогнозирования возможных негативных последствий, специфичных для алгоритмов МО. Для этого зависимости из табл. 1 должны быть представлены в виде матрицы чувствительности S, состоящей из векторов-столбцов s_{nm}, вида:

$$S = \{s_{nm}\} = \{\begin{matrix} inm \\ cnm \end{matrix}\}, \; n = 1..7, \; m = 1..4. \qquad (6)$$

Тогда, если вектор соответствия требованиям будет записан в виде вектора-строки:

$$A = \{i1, \; c1, \; i2, \; c2... \; i7, \; c7\}, \qquad (7)$$

где in и cn – бинарные коэффициенты, отражающие несоответствие (1) или соответствие (0) требованиям к целостности и конфиденциальности n-й информационной компоненты СИИ (табл. 1), то выражение для оценки параметров, характеризующих возможные негативные последствия, примет вид:

$$D = \{d_1, \; d_2... \; d_4\} = AS, \qquad (8)$$

а значение интегрального показателя уровня рисков для вектора соответствия требованиям A рассчитывается путем свертки параметров d_i с соответствующими весовыми коэффициентами:

$$d_0 = \Sigma_{i=1}^{4} \; \nu_i \, d_i \, . \qquad (9)$$

Для вычисления вектора параметров D и интегрального показателя D_0 значения элементов матрицы s_{nm} должны быть заданы на шкале отношений, допускающей операцию сложения, с учетом разнородного характера негативных последствий, обусловленных несоответствием различным требованиям.

Анализ рисков нарушения конфиденциальности данных должен проводиться с учетом специфики конкретной СИИ. В то же время, для рисков нарушения функциональной корректности могут быть сформулированы некоторые общие закономерности. Структура рисков, обусловленных нарушением функциональной корректности, с учетом перечисленных выше особенностей СИИ и в разрезе интересов и приоритетов различных заинтересованных сторон представлена в табл. 2.

Наиболее опасные угрозы и, соответственно, наиболее важные характеристики и требования связаны с обеспечением безопасности жизни и здоровья людей, а также с предотвращением крупных инцидентов экологической безопасности [7, 8]. Так или иначе, для систем повышенной опасности (п. 1 в табл. 2) необходимо иметь гарантии того, что уровень формируемых ими угроз не превышает уровень, демонстрируемый квалифицированными людьми-операторами, выполняющими соответствующие задачи управления и обработки данных в ручном режиме.

Особый вид требований связан с предотвращением угроз информационной безопасности (ИБ) в отношении заинтересованных лиц, вызванных нарушением функциональной корректности СИИ (п. 2 в табл. 2). Если некорректная работа систем может привести к реализации деструктивных информационно-психологических воздействий на общество (дезинформация, злонамеренное нарушение социальной стабильности), то в формировании требований ИБ к таким СИИ заинтересованы общество в целом и соответствующие государственные регуляторы (п. 2.2 в табл. 2).

Для многих прикладных СИИ специфичны угрозы этического характера и другие угрозы, предотвращение которых достигается реализацией мер т.н. «мягкого» права (п. 3 в табл. 2). К таким СИИ относятся, например, системы в кредитно-финансовой сфере и в области образования, поисково-справочные, маркетинговые и иные информационные системы, использующие методы персонализации на основе ИИ [9, 10].

Таблица 2.

Приоритеты заинтересованных сторон в области предотвращения угроз, обусловленных несоответствием функциональных характеристик СИИ установленным требованиям

Вид угроз, обусловленных нарушением функциональной корректности СИИ	Категория заинтересованной стороны	
	Лица, непосредственно участвующие в создании и применении СИИ (акторы ИИ)	Третьи лица
1. Угрозы жизни и здоровью людей, экологические угрозы	1.1. Потребители, разработчики и поставщики (собственная безопасность, дополнительные требования гос. регуляторов)	1.2. Общество в целом и регуляторы (безопасность общества и окружающей среды)
2. Угрозы информационной безопасности в отношении заинтересованных сторон	Нет	2.2. Общество в целом и государственные регуляторы (защита персональных данных, предотвращение деструктивных информационно-психологических воздействий)
3. Нарушение этических и других норм «мягкого» права	Нет	3.2. Общество в целом (социальная приемлемость создания и применения СИИ)
4. Неопределенные потребительские свойства, не влияющие непосредственно на безопасность жизни и здоровья людей, экологическую безопасность	4.1. Потребители (функциональные характеристики, определяющие возможность применения СИИ по назначению), разработчики и поставщики (характеристики конкурентоспособности СИИ)	Нет

Для СИИ, не предназначенных непосредственно для решения задач в области безопасности и не представляющих угрозы для жизни, здоровья людей и окружающей природной среды (п. 4 в табл. 2), отклонение функциональных характеристик от установленных требований ограничивается ухудшением потребительских свойств систем и может интерпретироваться, как реализация угроз экономической безопасности акторов ИИ. В формировании требований и предотвращении соответствующих угроз в таком случае заинтересованы, прежде всего, разработчики, поставщики и потребители СИИ (п. 4.1 в табл. 2).

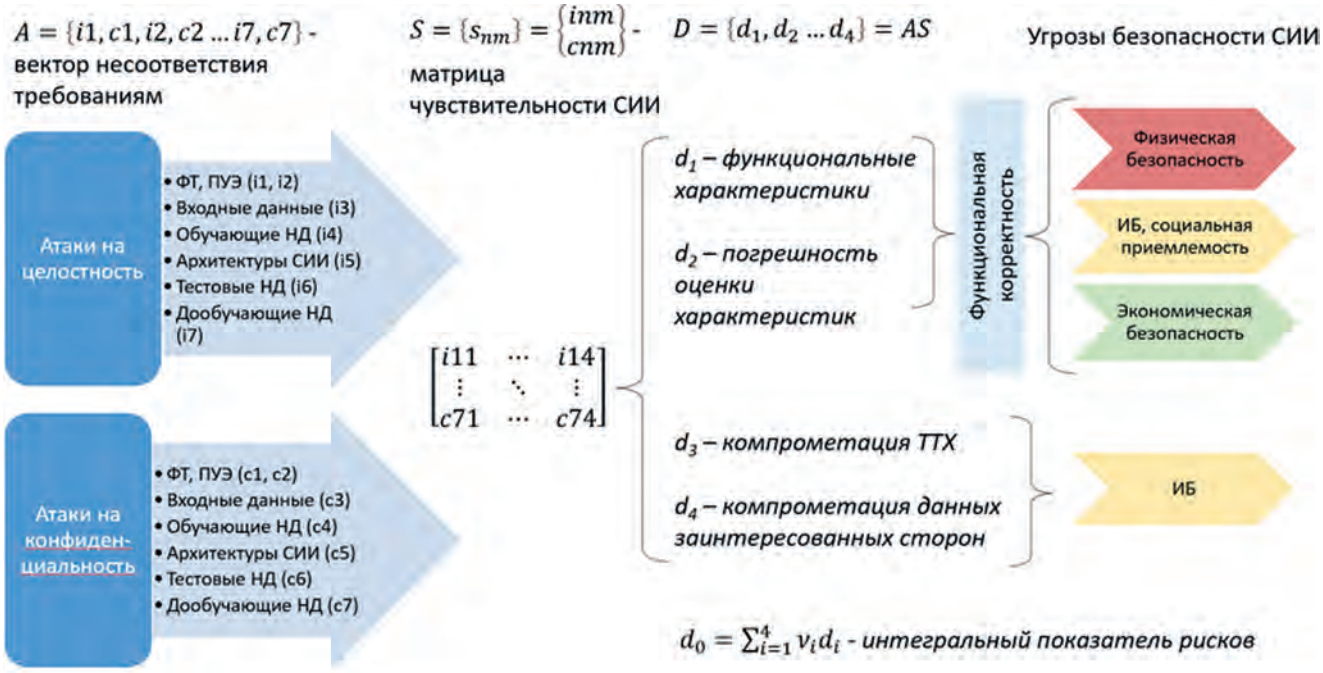


Рис.2. Специальная модель безопасности СИИ

Схема, поясняющая предложенную специальную модель безопасности СИИ, представлена на рис. 2. На основе данных о имеющихся и потенциальных несоответствиях требованиям в области целостности и конфиденциальности различных информационных компонент СИИ (A) и матрицы чувствительности СИИ к этим несоответствиям (S) оценивается вектор возможных негативных последствий (D), содержащий параметры функциональной корректности (d_1 и d_2) СИИ и конфиденциальности данных о СИИ и заинтересованных сторонах (d_3 и d_4, соответственно). Параметры функциональной корректности далее подвергаются дополнительному анализу с учетом прикладного значения и особенностей применения СИИ для оценивания рисков в области физической, ментальной (ИБ и социальная приемлемость) и экономической безопасности, после чего полученные оценки в сочетании с ранее полученными оценками рисков ИБ используются в качестве выходных данных модели. При необходимости может быть рассчитан показатель интегральных рисков, связанных с несоответствием СИИ установленным требованиям (d_0).

Предложенная модель построена с учетом общих особенностей систем обработки данных на основе алгоритмов машинного обучения, является качественной и иллюстрирует структуру зависимостей различного рода рисков, обусловленных несоответствием информационных компонентов СИИ различным требованиям. Количественное описание этих зависимостей представляет предмет дальнейших исследований с учетом отраслевой специфики прикладных систем ИИ.

Заключение

Таким образом, в статье рассмотрено влияние несоответствия различного рода требованиям, предъявляемым к информационным компонентам систем ИИ, на риски, возникающие при создании и применении этих систем. Анализ структуры зависимостей рисков с учетом общих особенностей систем обработки данных на основе алгоритмов машинного обучения показал, что эти риски сводятся к ухудшению функциональных характеристик СИИ, увеличению погрешности оценки этих характеристик эксплуатантом и компрометации данных, обрабатываемых в системе, причем часть рисков в области конфиденциальности данных представляет непосредственную угрозу, приводя к компрометации сведений об особенностях систем и о связанных с ними заинтересованных сторонах, а часть – создает дополнительные предпосылки для снижения функциональной корректности СИИ. Предложенная модель безопасности может быть использована для качественной оценки рисков, возникающих при создании и эксплуатации прикладных систем ИИ различного назначения, и организации мер по снижению этих рисков.

Литература

1. *Гарбук С. В., Губинский А. М. Искусственный интеллект в ведущих странах мира: стратегии развития и военное применение. – М.: Знание, 2020. 590 с.*
2. *Гарбук С. В. Метод оценки влияния параметров стандартизации на эффективность создания и применения систем искусственного интеллекта // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 3. С. 4–14.*
3. *Garbuk S. V. Intellimetry as a way to ensure AI trustworthiness // The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). Limassol, Cyprus, 6-10.10.2018. pp. 27–30.*
4. *Войнов Д. М., Ковалев В. А. Экспериментальная оценка состязательных атак на глубокие нейронные сети при решении задач распознавания медицинских изображений // Информатика., 2019. Т. 16. №3. С. 14–22.*
5. *Gary McGraw, Richie Bonett, Harold Figueroa, Victor Shepardson. Security Engineering for Machine Learning. Computer. IEEE Computer Society, 2019, vol.52, no. 8. pp. 54-57.*
6. *Унифицированная программная платформа машинного обучения «Платформа-ГНС» [Электронный ресурс] // Сайт ГосНИИАС. URL: https://www.gosniias.ru/platform.html (дата обращения: 01.09.2023).*
7. *Patrick Hall, James Curtis, Parul Pandey. Machine Learning for High-Risk Applications. Approaches for Responsible AI. 2023 April. 470 p.*
8. *Rahman, M. M. Should I Be Scared of Artificial Intelligence? // Academia Letters, Article 2536. DOI: https://doi.org/10.20935/AL2536.*
9. *O'Keefe K., Daragh O Brien. Ethical Data and Information Management: Concepts, Tools and Methods, Kogan Page. 2018. pp. 46-47, 214–218, 262-263.*
10. *Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. №2. С.84–105.*

АТАКИ И МЕТОДЫ ЗАЩИТЫ В СИСТЕМАХ МАШИННОГО ОБУЧЕНИЯ: АНАЛИЗ СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ

Котенко И. В.[1], Саенко И. Б.[2], Лаута О. С.[3], Васильев Н. А.[4], Садовников В. Е.[5]

DOI: 10.21681/2311-2024-1-24-37

Цель исследования: *проведение анализа атак на системы машинного обучения и методов защиты от них на основе известных обзорных статей, опубликованных за последние пять лет в высокорейтинговых журналах.*

Методы исследования: *системный анализ, классификация, моделирование, машинное обучение.*

Полученные результаты: *исследованы обзорные работы в высокорейтинговых журналах, посвященные анализу атак на системы машинного обучения и методам защиты от них. Выявлено, что тематика защиты от таких атак вызывает в настоящее время постоянно растущий интерес, а сфера воздействия таких атак охватывает интеллектуальные системы различного предназначения с ориентацией на широкий спектр типов обрабатываемых данных (изображения, звук, текст, видео, кибербезопасность и т.д.). Обобщены признаки классификации атак на системы машинного обучения и мер защиты от них. Выделены и рассмотрены наиболее распространенные атаки, которые по своему типу относятся к атакам «белого ящика» или «черного ящика». Обоснованы наиболее распространенные методы защиты от атак на системы машинного обучения, и дана их характеристика. Для ряда наиболее сложных методов защиты приведено их детальное описание на уровне отдельных этапов. Выделены особенности реализации методов защиты, позволяющие повысить эффективность обнаружения атак на системы машинного обучения.*

Научная новизна: *анализ работ по тематике атак на системы машинного обучения и мер защиты от них показал, что в настоящее время для них не существует устоявшейся классификации, что обусловлено бурным ростом новых разновидностей атак и появлением новых методов и механизмов защиты. Предложенные в рассмотренном исследовании признаки классификации атак и методов защиты обобщают подходы к такой классификации. Описание наиболее распространенных методов защиты отличается от других известных описаний поэтапной детализацией, которая обеспечивает простоту реализации этих методов в системах защиты интеллектуальных системах различного назначения.*

Вклад: *Котенко И. В. и Саенко И. Б. – общая концепция анализа атак на системы машинного обучения и методов защиты от них на основе известных обзорных работ; Котенко И. В. и Лаута О. С. – классификация и характеристика атак; Васильев Н. А. и Садовников В. Е. – классификация и поэтапная детализация мер защиты; Котенко И. В. и Саенко И. Б. – обсуждение особенностей реализации методов защиты.*

Ключевые слова: *кибербезопасность, машинное обучение, глубокое обучение, состязательные атаки, защита от атак, искусственный интеллект.*

ATTACKS AND DEFENSE METHODS IN MACHINE LEARNING SYSTEMS: ANALYSIS OF MODERN RESEARCH

Igor Kotenko[6], Igor Saenko[7], Oleg Lauta[8], Nikita Vasiliev[9], Vladimir Sadivnikov[10]

The purpose of the study: *conducting an analysis of attacks on machine learning systems and methods of protection against them based on well-known review works published in recent years in highly rated journals.*

Research methods: *system analysis, classification, modeling, machine learning.*

1 Котенко Игорь Витальевич, заслуженный деятель науки РФ, доктор технических наук, профессор, главный научный сотрудник и руководитель лаборатории проблем компьютерной безопасности, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ivkote@comsec.spb.ru

2 Саенко Игорь Борисович, доктор технических наук, профессор, ведущий научный сотрудник, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ibsaen@comsec.spb.ru

3 Лаута Олег Сергеевич, доктор технических наук, доцент, Государственный университет морского и речного флота им. адмирала С. О. Макарова (ГУМРФ), г. Санкт-Петербург, Россия. E-mail: laos-82@yandex.ru

4 Васильев Никита Алексеевич, научный сотрудник, Военная академия связи им. Маршала Советского Союза С.М. Будённого, г. Санкт-Петербург, Россия. E-mail: vasn2020@mail.ru

5 Садовников Владимир Евгеньевич, аспирант, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: bladimir1998@mail.ru

6 Igor Kotenko, Honored Worker of Science of the Russian Federation, Doctor of Technical Sciences, Professor, Chief Scientist and Head of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ivkote@comsec.spb.ru

7 Igor Saenko, Doctor of Technical Sciences, Professor, Leading researcher of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ibsaen@comsec.spb.ru

8 Oleg Lauta, Doctor of Technical Sciences, associate professor, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia. E-mail: laos-82@yandex.ru

9 Nikita Vasiliev, researcher, Military Telecommunication Academy named after the Soviet Union Marshal Budienny S.M., St. Petersburg, Russia. E-mail: vasn2020@mail.ru

10 Vladimir Sadovnikov, graduate student of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: bladimir1998@mail.ru

Results obtained: *review papers in high-ranking journals devoted to the analysis of attacks against machine learning systems and methods of protection against them were studied. It was revealed that the topic of protection against such attacks is currently of growing interest, and the scope of influence of such attacks covers intelligent systems for various purposes, focusing on a wide range of types of processed data (images, sound, text, video, etc.). The characteristics of the classification of attacks on machine learning systems and measures to protect against them were summarized. The most common attacks, which in their type belong to «white box» or «black box» attacks, were identified and considered. The most common methods of protection against attacks against machine learning systems are substantiated and their characteristics were given. For the most complex protection methods, their detailed description was given at the level of individual stages. Features of the implementation of protection methods were highlighted that make it possible to increase the efficiency of detecting attacks against machine learning systems.*

Scientific novelty: *an analysis of works on the topic of attacks against machine learning systems and measures to protect against them has shown that currently there is no established classification for them, which is due to the rapid growth of new types of attacks and the emergence of new protection methods and mechanisms. The characteristics of the classification of attacks and defense methods presented in the work generalize the approaches to such classification proposed in the works studied. The description of the most common protection methods differs from other known descriptions in its step-by-step detailing, which ensures ease of implementation of these methods in protection systems for intelligent systems for various purposes.*

Contribution: *Igor Kotenko and Igor Saenko – general concept of analysis of attacks against machine learning systems and methods of protection against them based on well-known review works; Igor Kotenko and Oleg Lauta – classification and characteristics of attacks; Nikita Vasilev and Vladimir Sadovnikov – classification and step-by-step detailing of protection measures; Igor Kotenko and Igor Saenko – discussion of the features of implementing protection methods.*

Keywords: *cyber security, machine learning, deep learning, adversarial attacks, attack protection, artificial intelligence.*

Введение

Искусственный интеллект (ИИ) и машинное обучение (МО) являются областями компьютерной науки, которые все шире используются для создания систем поддержки принятия решений, способных обрабатывать и анализировать информацию таким же образом, как это делает человек. Ключевой идеей функционирования систем МО является то, что они способны самостоятельно обучаться, анализируя большие объемы опытных данных, а затем использовать полученные знания для принятия решений в новых ситуациях. Эффективность систем МО обусловлена их способностью намного быстрее, чем человек, находить скрытые закономерности в данных, описывающих различные процессы или явления. Благодаря этому системы МО получают все более широкое внедрение во многих сферах деятельности человека [1].

В медицине системы МО используются для обработки медицинских данных, диагностики заболеваний и разработки индивидуальных планов лечения [1, 2]. В финансовой сфере системы МО применяются для прогнозирования рыночных трендов, оптимизации инвестиционных стратегий и улучшения управления рисками [4]. В промышленности системы МО применяются для автоматизации процессов, оптимизации планирования и управления качеством продукции [5]. На транспорте системы МО используются для оптимизации маршрутов и планирования логистики, улучшения безопасности на дорогах и разработки автономных транспортных средств [6]. В энергетике использование методов МО позволяет оптимизировать производство и планирование ресурсов, прогнозировать возможные перегрузки в электрической сети и принимать меры по более эффективному управлению энергией [7]. В сфере образования системы МО применяются для разработки адаптивных образовательных платформ, индивидуализации обучения, оптимизации пути обучения, выработки рекомендаций для улучшения результатов обучения [8].

Атаки на системы МО, иначе называемые состязательными (adversarial), могут приводить к нарушению безопасности и надежности функционирования целевых систем. Большинство состязательных атак сводится к изменению обучающего набора данных МО или изменению самой модели МО. Это может вызвать искажение результатов, полученных с применением технологии МО, что в критических системах приведет к серьезным последствиям и большим экономическим убыткам. Поэтому защита систем МО от атак становится в настоящее время одной из важнейших задач в области информационной безопасности и находит отражение в большом количестве научных трудов, посвященных этой тематике.

Количество работ, в которых рассматриваются состязательные атаки и методы защиты от них, возрастает с каждым годом. При этом каждый год появляются работы, в которых уточняются алгоритмы реализации таких атак и рассматриваются новые решения по противодействию этим атакам. Авторами настоящей работы также предлагались различные подходы и решения по защите от атак на системы МО [9–11].

В то же время представляет несомненный интерес проведение глубокого анализа известных атак и методов защиты в системах машинного обучения. Для этой цели авторы проанализировали обзорные работы по состязательным атакам и методам защиты от них, появившиеся в последние годы. Всего было отобрано для анализа свыше 30 обзорных работ, опубликованных за последние пять лет. Рассмотрение этих работ позволило выделить основные тренды в совершенствовании состязательных атак и методов защиты.

Целью настоящей работы является анализ атак на системы МО и методы защиты от них, проведенный на основе рассмотрения обзорных статей, посвященных состязательным атакам. Новизна работы заключается в представлении основных результатов и трендов в области реализации атак на системы МО и защиты от таких атак в рамках одной работы. В статье дается характеристика наиболее используемых состязательных атак и методов защиты. Рассмотрение методов защиты детализировано до алгоритмического уровня. Это позволяет данной статье помочь исследователям в области информационной безопасности систем МО.

Анализ работ по тематике атак на системы МО и методов защиты от них

Отбор работ для анализа осуществлялся следующим образом. Были задействованы поисковые системы Google.ru и Scholar.google.com (Академия Google). Использовалась строка поиска «Adversarial attack defense review». Период поиска был установлен в диапазоне 2020–2023 годы. Отбору подлежали обзорные статьи, опубликованные в журналах, имеющих квартиль Q1 – Q3 в международной индексной базе Scopus. Квартиль определялся по данным, представленным в системе https://www.scimagojr.com. Всего таким образом было отобрано для анализа 24 обзорные статьи. В таблице 1 приведены общие сведения об отобранных статьях.

Таблица 1

Общие сведения об отобранных обзорных статьях

П1	П2	П3	П4	П5	П6
[12]	2020	Q1	203	11	Глубокое обучение
[13]	2020	Q1	72	18	Вредоносный код
[14]	2021	Q2	139	25	Машинное обучение
[15]	2020	Q2	136	12	Изображения, графика, текст
[16]	2020	Q3	115	4	Глубокое обучение
[17]	2020	Q3	13	4	Электроэнергия
[18]	2021	Q1	450	339	Компьютерное зрение
[19]	2021	Q1	152	34	Умные грид-сети
[20]	2021	Q1	78	1	Глубокое обучение
[21]	2021	Q2	163	38	Кибербезопасность
[22]	2021	Q2	65	42	Умные грид-сети
[23]	2021	Q3	132	75	Изображения, текст, вредоносный код
[24]	2022	Q1	185	121	Глубокое обучение
[25]	2022	Q1	128	45	Глубокое обучение
[26]	2022	Q1	52	21	Глубокое обучение
[27]	2022	Q1	46	26	Цифровые сигналы
[28]	2022	Q1	34	27	Глубокое обучение
[29]	2022	Q2	103	48	Глубокое обучение
[30]	2022	Q2	49	21	Умные грид-сети
[31]	2022	Q2	46	7	Изображения
[32]	2023	Q1	246	140	Автономный транспорт
[33]	2023	Q1	176	119	Текст
[34]	2023	Q1	53	23	Кибербезопасность
[35]	2023	Q2	179	48	Графика

В таблице 1 учитываются следующие показатели: П1 – ссылка; П2 – год публикации; П3 – квартиль по базе Scopus; П4 – общее количество ссылок в статье; П5 – количество ссылок на источники, опубликованные с 2019 года; П6 – предметная область.

Анализируя данные по показателю П3, можно сделать вывод, что выбранные работы по тематике атак на системы МО имеют большую значимость и вызывают несомненный интерес. Подавляющее большинство работ было опубликовано в высокорейтинговых журналах квартилей Q1 и Q2.

Показатели П4 и П5 характеризуют широту охвата выбранных работ. В журналах квартилей Q1 и Q2, как правило, показатель П4 принимает значение, большее чем 100 (хотя есть несколько работ со значением гораздо ниже). Показатель П5, по нашему мнению, характеризует актуальность обзоров. Для работ, опубликованных в 2020 году, показатель П5 имеет небольшое значение, не превышающее 20, что является вполне естественным. Для следующих годов этот показатель увеличивается, принимая наивысшие значения для 2023 года. Это говорит о том, что, как правило, в различных обзорных работах можно найти ссылки на одни и те же исходные публикации, в которых рассматриваются конкретные случаи реализации различных типов состязательных атак и методов защиты.

Анализ показателя П6 позволяет сделать вывод, что большинство обзоров имеют широкую предметную область, которая отмечена как «глубокое обучение» или «машинное обучение» (9 обзоров). В то же время имеются обзоры, посвященные конкретной предметной области. В качестве таких предметных областей рассматриваются: изображения, графика, текст и/или вредоносный код – 6 работ; электроэнергия и умные грид-сети – 4; кибербезопасность – 2; компьютерное зрение – 1; цифровые сигналы – 1; автономный транспорт – 1. Такой широкий разброс предметных областей в обзорных статьях говорит о том, что тематика состязательных атак и защиты от них затрагивает в настоящее время практически все современные интеллектуальные системы.

В каждой из обзорных работ рассматривались классификация атак и мер защиты, а также давались характеристика и примеры их реализации, взятые из оригинальных источников. При этом следует отметить, что классификация атак и мер защиты из года в год уточнялась.

Далее рассмотрим наиболее распространенные атаки на системы машинного обучения и методы защиты от них, выделенные в результате анализа отобранных статей.

Атаки на системы МО

Для классификации атак на системы МО в работах [12–35] используются различные признаки, в частности, следующие:

- ✓ метод атаки (атаки на части модели МО или прямые атаки на данные);
- ✓ область данных, на которые направлена атака (изображения, звук, текст и т.д.);
- ✓ цель атаки (компрометация модели, уклонение от детектирования и т.д.);
- ✓ тип входных данные (непрерывные или дискретные);
- ✓ владение знаниями об атакуемой системе («белый ящик», «черный ящик», «серый ящик») и другие признаки.

Последний признак использовался в системах классификации атак, предлагаемых практически во всех работах. Поэтому в настоящей работе остановимся на рассмотрении этого признака.

Значение «белый ящик» означает, что злоумышленник полностью владеет необходимыми знаниями и о модели МО, и об обучающих наборах данных. «Черный ящик» – противоположный случай, когда у атакующего нет информации ни о модели МО, ни о наборах данных. Промежуточным вариантом является «серый ящик», когда злоумышленник владеет частичными, неполными знаниями о модели МО и наборах данных.

В таблице 2 представлено распределение наиболее известных состязательных атак по значениям признака владения знаниями. Название атаки дается в оригинале на английском языке, затем следует перевод и принятое для этой атаки сокращение.

Рассмотрим более подробно наиболее популярные методы атаки на системы МО.

FGSM (быстрый метод, основанный на знаке градиента) – это метод атаки «белого ящика» на нейронные сети, который используется для обмана моделей, обученных для распознавания изображений. Метод FGSM заключается в том, чтобы изменить изображение незначительно таким образом, чтобы обученная модель ошибочно идентифицировала его другим классом. Для этого используется метод градиентного спуска, который позволяет найти наиболее чувствительные пиксели на изображении. При использовании метода FGSM начальное изображение рассматривается как точка на пути от исходного до измененного изображения, которое обеспечивает максимальное изменение значения скорости потерь целевой функции (функции потерь). Затем производится вычисление градиента потерь по отношению к каждому пикселю изображения, после чего все пиксели с наименьшим модулем градиента

Таблица 2

Распределение атак по значениям признака владения знаниями

Значение признака	Название атаки (англ.)	Перевод названия	Сокращение
Белый ящик	Fast Gradient Sign Method	Метод быстрого градиента	FGSM
	Iterative Gradient Sign Method	Итеративный градиентный метод	IGSM
	Jacobian Saliency Map Attack	Атака карты значимости на основе Якобиана	JSMF
	Basic Iterative Method	Базовый итеративный метод	BIM
	Undetectable Perturbation	Незначительные изменения	UP
	Carlini and Wagner's Attack	Атака Карлини и Вагнера	C&W
	Iterative Least-Likely Class Method	Итеративный метод класса с наименьшей вероятностью	ILCM
	One-Step Target Class Method	Метод одношагового целевого класса	OSTCM
	Deep Fool	«Полный дурак»	DF
	Hot/Cold method	Горячий/холодный метод	HCM
	Ground-Truth Attack	Истина о системе	GTA
Черный ящик	Boundary Attack	Граничная атака	BA
	Zero-Query Attacks	Атака с нулевым запросом	ZQA
	Generative Adversarial Network	Генеративно-состязательная сеть	GAN
	One Pixel Attack	Атака одним пикселем	OPA
	Zeroth Order Optimization	Оптимизация нулевого порядка	ZOO
	Genetic Algorithms	Генетические алгоритмы	GA
	Improved Genetic Algorithm	Улучшенный генетический алгоритм	IFA
	Probability Weighted Word Saliency	Вероятностно-взвешенная значимость слова	PWWS
	Greedy Search Algorithm	Жадный алгоритм поиска	GSA
	Natural Evolution Strategies	Естественные эволюционные стратегии	NES
	Insertion and Removal of Words	Вставка и удаление слов	IRW
	Real-World Noise	Шум реального мира	RWN
Серый ящик	Cross-Site Scripting	Межсайтовый скриптинг	CSS
	Password Guessing	Подбор паролей	PG
	Buffer Overflow Attack	Атака переполнения буфера	BOA
	SQL Injection	SQL-инъекция	SQLI
	Weak Authentication Attack	Атака слабой аутентификации	WAA
	Cross-Site Request Forgery	Межсайтовая подделка запроса	CSRF

обнуляются, а остальные увеличиваются или уменьшаются на значение, которое составляет знак градиента. Метод FGSM позволяет создавать поддельные изображения, которые выглядят практически также, как оригиналы, но несут с собой измененную информацию, которая может обмануть модель машинного обучения. Существует несколько вариантов FGSM, которые отличаются тем, как определяется величина шага градиентного спуска. Например, FGSM может использоваться с фиксированным шагом либо определять шаг в каждой точке с применением линейного поиска с обратным ходом (backtracking line search). У FGSM есть ограничения. Одно из них заключается в том, что метод может обмануть модель только до определенной степени, после чего результаты перестают быть достоверными, и модель начинает идентифицировать измененное изображение правильно. Кроме того, FGSM может быть применен только к моделям, которые используют градиентный спуск для обучения. Следует также отметить, что методы атаки, такие как FGSM, могут быть использованы не только злоумышленниками, но и для различных исследовательских задач, связанных с оценкой уровня защиты нейронных сетей и их поведения в различных сценариях. В частности, метод FGSM может быть использован для разработки новых алгоритмов защиты нейронных сетей, позволяющих повышать уровень защиты от подобных атак.

IGSM (итеративный метод, основанный на знаке градиента) является разновидностью метода атаки «белого ящика» на нейронные сети, который расширяет возможности схожего алгоритма FGSM. Он основан на многократном применении метода FGSM с учетом нескольких изменений. IGSM является алгоритмом оптимизации, который начинается с исходного изображения и продолжает обновлять его через серию итераций с использованием FGSM. В каждой итерации значения пикселей изменяются в направлении увеличения потерь целевой функции. В отличие от FGSM, который использует только одну итерацию для создания поддельных изображений, IGSM повторяет процедуру атаки на каждой итерации, что дает лучший эффект, но требует больших вычислительных ресурсов. IGSM может быть использован как для целевой, так и для нецелевой атаки.

JSMA (атака карты значимости на основе Якобиана) – это алгоритм атаки «белого ящика» на системы определения поддельных изображений, основанный на методах глубокого обучения. Этот алгоритм использует вектор градиентов (Якобиан), который определяет, как изменения весов в нейронной сети повлияют на выходной результат. В результате JSMA может определить наиболее «важные» признаки (части) изображения, которые влияют на классификацию модели. JSMA начинается с выбора целевой модели для атаки. Затем вычисляется вектор градиентов для каждой части изображения, позволяющий определить те части, которые влияют на классификацию картинки как подделку. Затем увеличивается влияние этих частей изображения, уменьшая влияние других частей, и приводя к результату, когда нейронная сеть классифицирует подделку.

BIM (базовый итеративный метод) – это тип атаки «белого ящика» на системы МО, который основан на внедрении изменений во входные данные. Результаты атаки могут привести к ошибочным выводам или неправильным действиям системы. Примеры BIM-атак включают изменение значений входных параметров при обучении моделей машинного обучения. Например, если система обучается определять различные образцы на основе цвета, размера и формы, то злоумышленник может ее обмануть, предоставив входные данные, содержащие измененные значения цвета, размера и формы. Другой пример BIM-атаки может быть направлен на автоматизированные системы контроля качества, когда злоумышленник отправляет измененные данные, создавая ложные сигналы об ошибке. Такие атаки могут вызвать сбои в системе, неправильную работу оборудования или опасные сбои в производственном процессе. Для предотвращения BIM-атак необходимо включать меры безопасности при разработке и настройке систем МО, такие как проверка входных данных, использование контроля целостности данных и обучение моделей на большом количестве данных. Также следует использовать методы дополнительного контроля, такие как применение одноразовых ПИН-кодов или двухфакторной аутентификации.

UP-атака (введение незначительных изменений) – это тип атаки «белого ящика», который заключается во внедрении незначительных изменений в данные или параметры модели, приводящих к ошибочным выводам и дискредитации результатов. Основная цель такой атаки – обойти систему защиты и создать искаженные данные, чтобы они были приняты за правильные. UP-атака может быть использована в различных областях, например, для манипулирования результатами голосования, изменения прогнозов погоды, машинного обучения, автономных транспортных средств, медицинских диагностических систем. Она является достаточно сложной для обнаружения, так как создает незначительные изменения в данных.

BA (граничная атака) – это типовой метод атаки «черного ящика», основанный на принятии решений. Начиная с исходного состязательного изображения, в нем используется бинарный поиск для нахождения точки выборки, которая находится вблизи границы классификации. Производится случайное блуждание по границе между двумя противоположными областями, чем уменьшается расстояние от целевого изображения. В соответствии с этим шагом продолжается итерация и постепенно уменьшается расстояние от исходного изображения. Причина, по которой этот тип алгоритма называется «граничной атакой», заключается в том, что он генерирует состязательные примеры путем поиска вдоль границы до тех пор, пока они не сойдутся для получения оптимального или рационального решения. Результаты, полученные таким методом, могут удовлетворять требованиям ошибочной классификации модели «черного ящика». Общее возмущение, которое увеличивается по сравнению с исходным изображением, зависит от производительности алгоритма.

ZQA (атака с нулевым запросом) – это атака «черного ящика», которая предназначена для передачи опыта между моделями без доступа к информации входных данных. В их основе лежит передача знаний между моделями, используя выводы моделей, а не входные данные. Традиционно для передачи опыта между моделями требуется доступ к исходным данным моделей, что может привести к утечке конфиденциальной информации. Злоумышленники могут применять атаки ZQA для выполнения различных задач, например, для создания фальшивых изображений и видео, которые позволяют обмануть системы компьютерного зрения, или для атак на защищенные системы распознавания лица, используя

данные полученные от других систем распознавания. Кроме того, атаки ZQA могут быть использованы для идентификации конфиденциальной информации. Например, злоумышленники могут использовать их для обнаружения ключевых слов и фраз в документах, которые не должны быть доступны публично. Они могут использовать знания, полученные от одной модели, чтобы обучить другую модель, которая может идентифицировать эти конфиденциальные данные. В то же время, атаки ZQA могут быть использованы и в благих целях. Например, они могут использоваться для передачи опыта между моделями в области медицины или научных исследований, позволяя ускорить процесс обучения модели и позволить получить более точные результаты. Защита от ZQA может быть построена на основе использования методов обнаружения аномалий и обучения с учителем. Эти методы могут идентифицировать необычные выходные данные, которые могут быть связаны с ZQA. Другим возможным направлением защиты является разработка методов обнаружения и предотвращения передачи опыта между моделями, используя только выходные данные. Защиту можно усилить с помощью обучения моделей для препятствования реализации атак и замедлять процесс передачи опыта между моделями.

Атака с использованием GAN (генеративно-состязательной сети) – это метод атаки «черного ящика», использующий нейронные сети GAN для формирования различных атак на модели МО. Принцип работы GAN заключается в тренировке двух нейронных сетей – генератора и дискриминатора, которые последовательно передают друг к другу данные и обучаются. На первом этапе генератор создает поддельные примеры данных, которые передаются дискриминатору вместе с настоящими примерами из обучающего набора. Дискриминатор обучается отличать настоящие данные от поддельных, и генератор учится создавать такие данные, чтобы их было сложно отличить от реальных. На втором этапе генератор использует полученные знания о структуре данных, чтобы создать атаки на модель машинного обучения. Эти атаки могут быть различными в зависимости от типа модели и задачи, которую она решает. Как только атака сгенерирована, она может быть использована злоумышленником для нападения на модель МО. Таким образом, GAN позволяет генерировать различные виды атак на модели МО, что делает их более уязвимыми для нападений. Это может быть использовано для тестирования устойчивости моделей и нахождения уязвимостей в их защите.

OPA (атака одним пикселем) относится к атакам «черного ящика» и основывается на алгоритмах МО. Она использует уязвимости в работе нейронных сетей, которые определяют изображения на основе цветовых значений каждого пикселя. Основной принцип работы этой атаки состоит в том, чтобы изменить значение всего лишь одного пикселя на изображении таким образом, чтобы нейронная сеть неправильно классифицировала это изображение. Например, при редактировании фото с котом, атака OPA может изменить значение пикселя на месте носа кота таким образом, что нейросеть будет считать, что на самом деле изображается собака. Атака OPA использует эволюционные алгоритмы, позволяющие определить оптимальные пиксели и изменить их значения таким образом, чтобы обмануть нейронную сеть. Использование таких алгоритмов позволяет достичь максимальной эффективности атаки при минимальном числе изменений на изображении.

NES (естественные эволюционные стратегии) – это семейство алгоритмов численной оптимизации для задач «черного ящика». Как и все другие эволюционные стратегии, они итеративно обновляют параметры поискового распределения, следуя естественному градиенту в сторону более высокой ожидаемой приспособленности. Общая процедура заключается в следующем. Для создания множества точек поиска используется параметризованное распределение. В каждой точке оценивается функция соответствия. Параметры распределения позволяют алгоритму адаптивно фиксировать значения функции приспособленности. Например, в случае распределения Гаусса они включают среднее значение и ковариационную матрицу. На основе выборок оценивается градиент поиска в сторону более высокой ожидаемой пригодности. Затем выполняется шаг подъема вдоль естественного градиента. Этот шаг имеет решающее значение, так как он предотвращает колебания, преждевременное схождение и нежелательные эффекты, возникающие из-за заданной параметризации. Весь процесс повторяется до тех пор, пока не будет выполнен критерий останова.

Методы защиты от атак на системы МО

Возможными признаками классификации методов защиты от атак на МО, указанными в работах [12–35], являются:

- ✓ направленность защиты (наборы данных, модель МО);
- ✓ способ анализа наборов данных (обнаружение изменений, защита от предобработки);
- ✓ способ обработки модели МО (обнаружение состязательных примеров, укрепление модели);
- ✓ направленность на слой нейронной сети (входной слой, промежуточные слои, выходной слои) и другие.

В результате можно выделить следующие наиболее популярные методы защиты (рис. 1):

1) состязательная тренировка (competitive training);
2) оборонительная дистилляция (defensive distillation);
3) реконструкция входных данных (input data reconstruction);
4) фреймворк Defense-GAN;
5) подкрепление (укрепление) модели (model reinforcement);
6) защита от предварительной обработки (protection from preprocessing);
7) обнаружение примеров состязательности (detection of adversarial examples).

Среди перечисленных методов только Defense-GAN реализуется с использованием соответствующего фреймворка. Остальные методы могут быть реализованы на основе применения различных средств.

Рассмотрим подробнее содержание этих методов защиты.

Рис. 1. Основные методы защиты от атак на системы МО

Состязательная тренировка. «Состязательная тренировка» – это метод защиты от атак и взломов систем МО, который использует GAN. Как было сказано выше, нейронная сеть GAN состоит из двух компонентов: генератора и дискриминатора. Генератор создает поддельные данные, а дискриминатор учится отличать их от настоящих. Генератор учится создавать данные таким образом, чтобы они были максимально похожи на настоящие. Дискриминатор учится отличать их от настоящих с большой точностью. Этот процесс продолжается до тех пор, пока генератор не научится создавать данные, которые дискриминатор не сможет отличить от настоящих. Таким образом, система МО, использующая состязательную тренировку, может обучиться отличать подделки от настоящих данных, что делает ее более защищенной от атак и взломов.

В процессе реализации метода состязательной тренировки можно выделить следующие этапы (рис. 2).

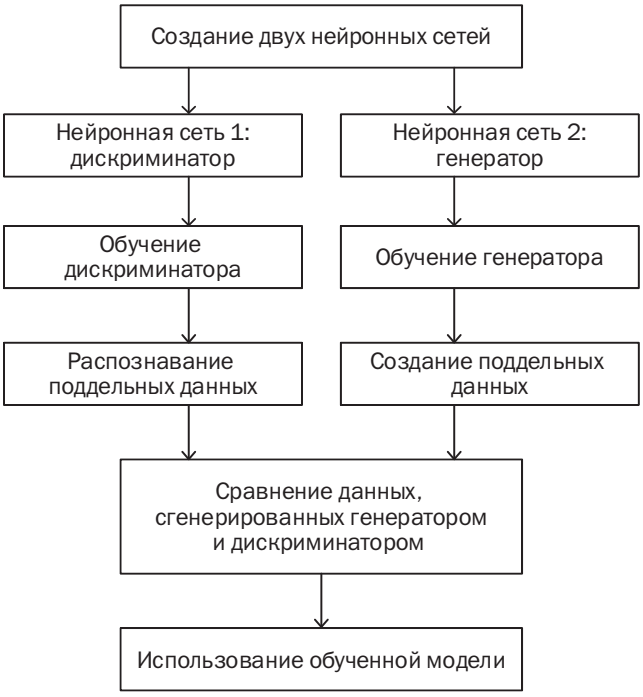


Рис. 2. Этапы реализации метода состязательной тренировки

Этап 1. Создание двух нейронных сетей: генератора и дискриминатора.

Этап 2. Обучение генератора созданию поддельных данных (изображений, звуков, текстов и т.д.). Параллельно дискриминатор обучается отличать поддельные данные от настоящих.

Этап 3. Генератор и дискриминатор конкурируют между собой в рамках задачи, которая была определена для тренировки модели (например, распознавание лиц).

Этап 4. Генератор создает новые поддельные данные, а дискриминатор оценивает, насколько они похожи на реальные данные. Оценка дискриминатора передается обратно генератору, чтобы тот мог улучшить свои навыки.

Этап 5. Дискриминатор обучается становиться все более точным в распознавании подделок. Генератор, в свою очередь, учится создавать все более качественные подделки. Этот процесс повторяется множество раз, пока подделки не станут практически неотличимыми от настоящих данных.

Этап 6. Модель, полученная после тренировки, может быть использована для анализа новых данных. Например, она может использоваться для распознавания лиц на фотографиях, прикрепленных к заявкам на кредит, для выявления лжи при беседе с клиентом и т.д.

В зависимости от приложений и областей, в которых применяется этот метод, шаги могут незначительно изменяться для достижения лучших результатов. Однако общий подход остается прежним и заключается в конкурентной тренировке генератора и дискриминатора с целью создания лучших поддельных данных и защиты систем МО от атак и взломов.

Оборонительная дистилляция. Оборонительная дистилляция – это метод защиты, базирующийся на создании и использовании так называемой «отфильтрованной» или «дистиллированной» версии данных, которую алгоритмы МО могут использовать для обучения своих моделей. Данные проходят через фильтры и алгоритмы, которые идентифицируют и удаляют такие элементы данных, которые могут быть опасными для системы. Например, это могут быть данные, содержащие вредоносный код или данные со специальными символами, которые предназначены для реализации атак на систему.

Оборонительная дистилляция, как правило, включает в себя несколько этапов: определение набора данных для обучения, анализ и фильтрацию данных, создание обучающих моделей и их эксплуатацию.

Набор данных, используемый для оборонительной дистилляции, должен быть репрезентативным и содержать данные, которые могут быть использованы для определения и обнаружения типов атак и уязвимостей в системе. Эти данные подвергаются анализу и фильтрации. После фильтрации данных они используются для обучения моделей МО, которые, в свою очередь, могут использоваться для обнаружения потенциальных угроз безопасности и принятия соответствующих мер.

Существенным преимуществом оборонительной дистилляции является ее способность к защите системы от новых видов атак. Традиционные методы защиты, такие как фильтрация трафика и использование антивирусов, направлены преимущественно на обнаружение и блокирование известных угроз, в то время как оборонительная дистилляция способна обнаружить и защитить систему от новых видов атак, которые еще не известны.

Реконструкция входных данных. Метод реконструкции входных данных основан на идее создания механизма защиты, который позволяет анализировать входные данные и осуществлять их реконструкцию, которая затем сравнивается с начальными входными данными. Если входные данные были изменены злоумышленником, то реконструкция будет отличаться от начальных данных, что позволит сигнализировать о возможности атаки на систему. Однако необходимо учитывать, что этот метод может иметь высокий уровень ложных срабатываний.

Процесс реализации метода реконструкции входных данных можно разделить на следующие этапы.

Этап 1. Обработка входных данных. На этом этапе входные данные проходят предварительную обработку, например, они могут быть преобразованы в числовой вид, аномалии и шум могут быть удалены.

Этап 2. Создание реконструкции на основании модели, используемой для обучения, работы системы МО и в соответствии с правилами обработки данных.

Этап 3. Сравнение реконструкции с изначальными входными данными. Если восстановленные данные отличаются от исходных данных, то предупреждение о том, что, возможно, система была атакована, входные данные были заменены или изменены. Другая причина – вероятность ошибки при выполнении алгоритма превышает разумный уровень.

Этап 4. Принятие соответствующих мер. Например, можно прекратить обучение или работу, попросить у пользователя подтверждение правильности входных данных или оповестить администратора о возможной атаке.

Важным аспектом работы метода реконструкции входных данных является выбор модели, используемой для создания реконструкции. Эта модель должна быть способна точно восстанавливать входные данные при минимальной потере информации. Выбор модели зависит от конкретной задачи и особенностей данных.

Defense-GAN. Defense-GAN – это фреймворк, предназначенный для защиты от атак на GAN-сети. Defense-GAN настраивает защищаемые модели путем генерации видоизмененных исходных данных, что делает атакующую модель недействительной, так как она обучается на искаженной информации.

Для определения эффективности искажений Defense-GAN использует статистические метрики, которые оценивают, насколько хорошо искажения защищают модель от атаки. Если метрики показывают, что защищаемая модель имеет хорошую защиту от атак, то генерируемые искажения можно использовать для защиты от нежелательных воздействий.

Кроме того, Defense-GAN использует критерии обучения, направленные на защиту, которые основаны на минимизации потерь при классификации и систематическом сдвиге наиболее важных признаков на изображении. Эти критерии обучения помогают защищаемой модели лучше предсказывать классы, а также улучшают ее устойчивость к атакам.

В процессе защиты с использованием фреймворка Defense-GAN можно выделить отдельные этапы (рис. 3).

Этап 1 (подготовка данных). Этот этап включает загрузку данных и их предварительную обработку. Для обучения фреймворка необходимы наборы

изображений, которые будут использоваться для обучения генератора и классификатора. Создание такого набора может включать в себя множество шагов предварительной обработки, таких как изменение размеров, повороты, зеркальные отражения и другие.

Этап 2 (обучение генератора). На этом этапе фреймворк обучается создавать защищенные версии исходных изображений на основе целевой функции, используя GAN-генератор. Генератор обучается создавать новые, но похожие на исходные изображения, которые будут менее чувствительны к различным типам атак.

Этап 3 (обучение дискриминатора). После обучения генератора начинается обучение дискриминатора – классификатора, который будет использоваться для оценки качества изображений. Классификатор обучается давать наиболее точную оценку классов, которым принадлежат изображения, а также определять, насколько защищены изображения.

Этап 4 (тестирование). После обучения модели проводится ее тестирование на тестовом наборе данных, который не использовался при обучении. Это позволяет оценить способность фреймворка защищать изображения от различных типов атак и определить, насколько точным является классификатор.

Этап 5 (защита от атак). При обнаружении атаки на модель глубокого обучения механизмы защиты могут включаться автоматически, используя защищенные версии изображений, созданные GAN-генератором. Это может помочь защитить модель от перебросок искажений, уменьшения качества изображений и других типов атак.

Этап 6 (оценка качества). Оценка качества модели осуществляется путем оценки ее способности защитить модель глубокого обучения от разных типов атак. При этом могут использоваться такие метрики, как достоверность (accuracy) и F-мера.

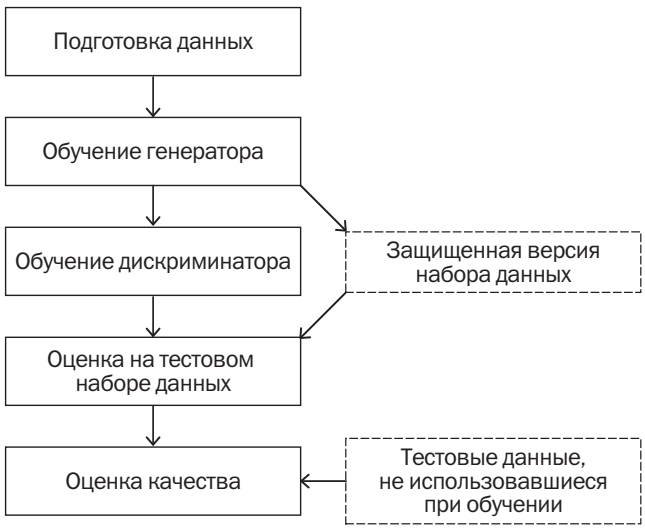


Рис. 3. Этапы работы фреймворка Defense-GAN

Подкрепление модели МО. Подкрепление модели МО – это метод защиты, который заключается в том, чтобы создать дополнительный слой защиты, добавляющий «укрепление» к системе, защищая ее от атак, ориентированных на модель. Идея заключается в использовании уже созданных моделей, которые определяют, какие действия являются безопасными, а какие – потенциально вредными. Система может использовать эти модели, чтобы определить, какие действия должны быть разрешены, и какие – запрещены. Если действие не соответствует модели безопасности, то оно будет заблокировано.

Один из примеров использования метода подкрепления модели – это алгоритм деревьев решений. Он использует деревья для принятия решений, основанных на определенных факторах безопасности. Если действия пользователя соответствуют модели, то алгоритм разрешит их. Если нет, то действия будут заблокированы.

Защита от предварительной обработки данных. Защита от предварительной обработки данных, также известная как «защита трансформаций», представляет собой метод защиты компьютерных систем от атак, основанных на предварительной обработке данных перед их отправкой на сервер. Метод предполагает изменение формы или контента сообщения перед их отправкой на сервер с целью затруднить анализ данных злоумышленником.

Этапы метода защиты от предварительной обработки данных (защиты трансформаций) представлены на рис. 4.

Этап 1 (анализ данных). Сначала компьютерная система анализирует данные, которые будут отправлены на сервер. Это могут быть, например, данные, введенные пользователем на веб-странице или в приложении.

Этап 2 (трансформация данных). Затем система трансформирует данные таким образом, чтобы изменить их форму или контент. Например, это может быть замена символов на другие символы или добавление случайных данных, чтобы затруднить распознавание оригинальных данных.

Этап 3 (отправка данных на сервер). После трансформации данные отправляются на сервер для дальнейшей обработки.

Этап 4 (распознавание трансформации). Когда данные приходят на сервер, система должна распознать трансформацию, которая была применена к данным.

Этап 5 (обработка данных). После распознавания трансформации система должна обработать данные с учетом трансформации. Это может включать в себя дешифрование данных или удаление случайной информации, которая была добавлена к данным.

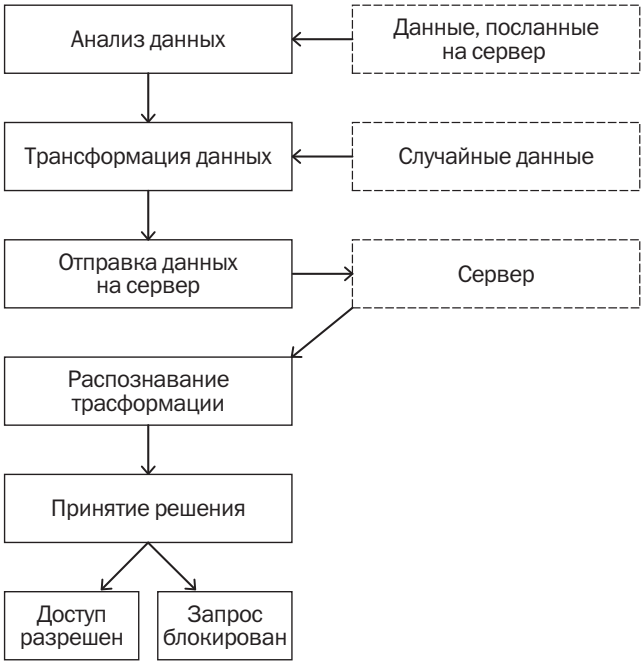


Рис. 4. Этапы метода защиты от предварительной обработки данных

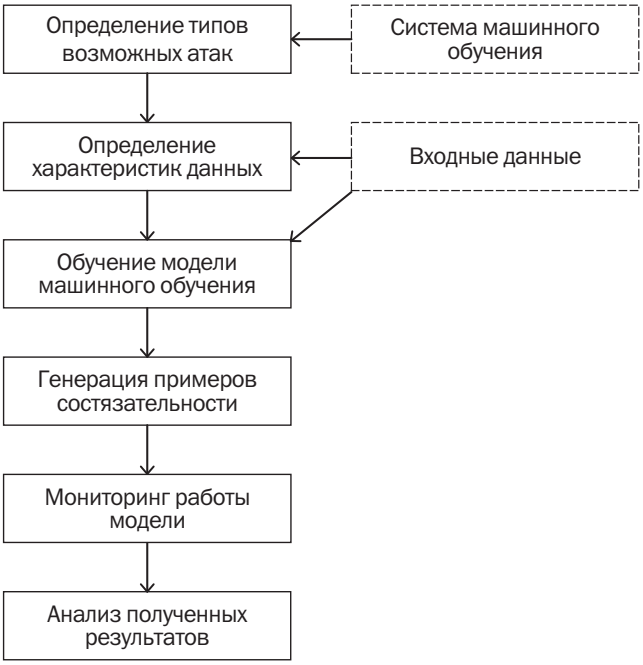


Рис.5. Этапы метода обнаружения примеров состязательности

Этап 6 (принятие решения). На основе обработанных данных система принимает решение, как обрабатывать запрос. Например, она может разрешить доступ к определенным ресурсам или заблокировать запрос, если он не соответствует политикам безопасности.

В целом, метод защиты от предварительной обработки данных может помочь защитить компьютерную систему от многих видов атак, которые основываются на предварительной обработке данных. Однако он не является единственным и должен использоваться в сочетании с другими методами.

Обнаружение примеров состязательности. Метод обнаружения примеров состязательности (Adversarial Examples) заключается в поиске и обнаружении таких примеров данных, которые могли бы использоваться для атак на модель МО. Примеры состязательности – это измененные данные, которые были подготовлены для того, чтобы обмануть модель МО. Например, это могут быть изображения, на которых были внесены незначительные изменения, которые затрудняют или делают невозможным их распознавание моделью.

Метод обнаружения примеров состязательности может быть реализован с использованием следующих подходов:

1) мониторинг необычных изменений в поведении модели МО; если модель начинает давать неправильные ответы или сильно изменяется ее точность, это может быть признаком того, что она подвергается атаке состязательных примеров;
2) анализ данных на предмет выявления изменений в распределении; примеры состязательности могут изменить распределение входных данных, что может быть замечено при анализе статистических показателей;
3) использование тестовых наборов данных с примерами состязательности для обучения модели; если модель обучена на данных с такими примерами, она может стать более устойчивой к таким атакам в будущем.

Метод может снижать свою точность при работе с обычными данными. Поэтому необходим баланс между защитой от примеров состязательности и сохранением высокой точности модели.

Процесс мониторинга необычных изменений в работе модели МО для обнаружения примеров состязательности делится на следующие этапы (рис. 5).

Этап 1 (определение типов возможных атак). Необходимо определить, какие типы примеров состязательности могут быть применены для атаки на модель МО, чтобы затем разработать методы обнаружения таких атак.

Этап 2 (определение характеристик данных). Следует провести анализ входных данных, на основе которых модель МО принимает решения, и выявить их наиболее значимые характеристики и признаки.

Этап 3 (обучение модели МО). Нужно обучить модель МО на основе данных, которые не содержат примеров состязательности, чтобы получить базовую версию модели.

Этап 4 (генерация примеров состязательности). Необходимо сгенерировать различные примеры

состязательности, которые могут быть использованы для атаки на модель МО, на основе знаний о характеристиках входных данных и типах возможных атак.

Этап 5 (мониторинг работы модели). Нужно наблюдать за изменениями в работе модели МО, которые могут свидетельствовать о наличии входных данных, содержащих примеры состязательности. Это может происходить на основе анализа метрик качества обучения, таких как точность, время обработки или показатели ошибок.

Этап 6 (анализ и документация результатов). Необходимо проанализировать полученные результаты и задокументировать эффективность методов обнаружения примеров состязательности, чтобы повысить эффективность системы защиты.

Проблема защиты систем машинного обучения от атак

Искусственный интеллект и, в частности, машинное/глубокое обучение являются мощным инструментом в сфере информационной безопасности. Эти технологии могут быть использованы как для защиты систем, так и для реализации атак. Применение МО может значительно улучшить методы защиты систем. МО и анализ данных позволяют разрабатывать более сложные и инновационные методы по обнаружению атак и предотвращению угроз.

Однако при использовании МО при реализации атак возникает серьезная проблема: системы защиты могут стать бессильными перед алгоритмами, использующими МО. Реализация атак на системы защиты, основанные на методах МО, представляет серьезную угрозу. Например, атакующая сторона может использовать нейронные сети для создания фальшивых данных и обмана системы защиты. Это приводит к неправильным решениям или проникновению в систему через механизмы защиты, которые не смогут идентифицировать подобные атаки.

Фреймворки, основанные на МО, способны выявлять аномальное поведение и реагировать на новые виды атак, которые ранее были неизвестны. Стоит заметить, что несмотря на все достоинства фреймворка Defense-GAN, у него существует следующий недостаток. Отсутствие зависимости от точки инициализации нейронной сети в прикладных задачах защиты информации влечет за собой то, что оптимальный дискриминатор будет присваивать более высокое значение для функции потерь, чем самим частям реальных обрабатываемых данных из генератора. Если модифицировать работу дискриминатора, то сам дискриминатор сразу же станет неоптимальным. По этой причине функционирование фреймворка Defense-GAN должно быть скомпоновано с соответствующей функцией потерь, которая будет учитывать описанную выше особенность для той предметной области знаний, в рамках которой решается задача обеспечения информационной безопасности.

Метод OPA весьма неэффективен, если в базовой конструкции модели машинного обучения используется два и более слоя пулинга. При этом неважно, какой тип пулинга используется, поскольку комбинация слоев данного типа нивелирует эффект разности пикселей на уровне высокоуровневых признаков.

Нужно признать, что метод оборонительной дистилляции имеет особенности при защите моделей ансамблей. Так, если в решающей модели будет использоваться алгоритм с привилегированной информацией, то раздельное функционирование модели-ученика и модели-учителя может привести к коллизиям в процессе нормальной работы базовой модели.

Метод JSMA также обладает следующей важной особенностью. Он не может функционировать одновременно с моделью МО. Поэтому правильная организация потоков данных в пайп-лайне построения модели МО поможет полностью исключить негативный эффект от внедрения JSMA в качестве вредоносного компонента. При этом не потребуется каких-либо дополнительных надстроек, контролирующих процесс функционирования основной модели МО.

Наконец, следует заметить, что защита от атак является весьма нетривиальной. Основная модель МО может функционировать неправильно при появлении в наборе данных аномальных величин. К аномальным показателям относятся не только попытки взлома, но и, чаще всего, просто некоторые редкие значения. Например, рост человека в 240 см может показаться аномальным для обобщающей способности алгоритма, если модель чаще всего оперировала с данными среднестатистического роста взрослого человека. По этой причине внедрять компоненты защиты от атак следует с особой осторожностью, так как нестандартные данные могут быть интерпретированы системой защиты как попытка несанкционированного доступа к данным.

Заключение

В статье представлены результаты анализа обзорных работ по реализации атак (состязательных атак) и методам защиты в системах машинного обучения, опубликованных за 2020–2023 годы в высокорейтинговых журналах. Анализ показал, что тематика защиты от состязательных атак вызывает в настоящее время постоянно растущий интерес и наблюдается бурный рост исследований в этой области. Причем эта сторона информационной безопасности касается различных видов обрабатываемых данных (изображения, звук, текст, компьютерное зрение и т.д.) и различных сфер жизни человека (медицина, транспорт, финансы, экономика и т.д.).

DOI: 10.21681/2311-3456-2024-1-24-37

В ходе анализа атак на системы МО выделены наиболее часто используемые в обзорах признаки их классификации и дана общая характеристика наиболее распространенных атак, которые по своему типу относятся к атакам «белого ящика», «черного ящика» и «серого ящика». В качестве наиболее распространенных атак «белого ящика» выделены атаки типов FGSM, IGSM, JSMF, BIM и UP. В качестве наиболее распространенных атак «черного ящика» выделены атаки типов BA, ZQA, GAN и OPA.

Выделены признаки классификации методов защиты от состязательных атак и дана характеристика наиболее распространенных методов. К числу таких методов были отнесены состязательная тренировка, оборонительная дистилляция, реконструкция входных данных, фреймворк Defense-GAN, подкрепление модели, защита от предварительной обработки и обнаружение примеров состязательности. Для наиболее сложных методов защиты приведено их детальное описание на уровне отдельных этапов.

Результаты проведенного анализа могут быть использованы для разработки новых эффективных методов и механизмов защиты от угроз, связанных с технологией МО. Например, для создания методики, способной оценивать и выбирать методы защиты систем МО, что в настоящее время является одной из ключевых проблем в сфере информационной безопасности.

Рецензент: *Липатников Валерий Алексеевич, доктор технических наук, профессор, научный сотрудник научно-исследовательского центра Военной академии связи имени Маршала Советского Союза С. М. Буденного, Санкт-Петербург, Россия. E-mail: lipatnikovanl@mail.ru*

Литература

1. *Клименко Р. В., Тараroев Я. В. Философское осмысление применения технологий машинного обучения. Перспективы искусственного интеллекта // Социальное время. 2016. № 1 (5). С. 15–30.*
2. *Понкин И. В. Цифровые модели-двойники пациентов: понятие и правовые аспекты // Бизнес, менеджмент и право. 2022. № 2 (54). С. 10–14.*
3. *Иванько А. Ф., Иванько М. А., Гаврилов К. А. IT-технологии обучения и их применение в различных сферах // Молодой ученый. 2019. № 1 (239). С. 5–10.*
4. *Пилявская И. М. Аналитический обзор применения технологий машинного обучения в финансовых ассистентах // Вестник науки и образования. 2022. № 4-2(124). С. 29–34. DOI: 10.24411/2312-8089-2022-10402.*
5. *Сааков Д. В. Применение методов машинного обучения для оптимизации производственных процессов в металлургической промышленности // Инновации и инвестиции. 2023. № 5. С. 308–311.*
6. *Проневич О. Б., Зайцев М. В. Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте // Надежность. 2021. Т. 21. № 3. С. 54–65. DOI: 10.21683/1729-2646-2021-21-3-54-65.*
7. *Энгель Е. А., Энгель Н. Е. Методы машинного обучения для задач прогнозирования и максимизации выработки электроэнергии солнечной электростанции // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2023. № 2. С. 146–170. DOI: 10.17308/sait/1995-5499/2023/2/146-170.*
8. *Мочалина М. В., Цапина Т. Н., Чайкина Ж.В . Использование машинного обучения в образовании // Russian Journal of Education and Psychology. 2023. Т. 14. № 1-2. С. 136–140.*
9. *Kotenko I., Saenko I., Lauta O., Kribel K., Vasiliev N. Attacks on artificial intelligence systems: classification, the threat model and the approach to protection // Proceedings of the Sixth International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'22). IITI 2022. Lecture Notes in Networks and Systems, vol 566. Springer, Cham. 2023. Pp. 293–302. DOI: 10.1007/978-3-031-19620-1_28.*
10. *Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks Against Machine Learning Systems: Analysis and GAN-based Approach to Protection // Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). IITI 2023. Lecture Notes in Networks and Systems, vol 777. Springer, Cham. 2023, pp. 49–59. DOI: 10.1007/978-3-031-43792-2_5.*
11. *Котенко И. В., Саенко И. Б., Лаута О. С., Васильев Н. А., Садовников В.Е. Подход к обнаружению атак на системы машинного обучения с использованием генеративно-состязательной сети // Двадцать первая Национальная конференция по искусственному интеллекту с международным участием, КИИ-2023 (Смоленск, 16-20 октября 2023 г.). Труды конференции. В 2-х томах. Т.1. 2023. С. 366–376.*
12. *Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability // Computer Science Review, Volume 37, 2020, 100270. DOI: 10.1016/j.cosrev.2020.100270.*
13. *Martins N., Cruz J. M., Cruz T., Henriques Abreu P., Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review // IEEE Access, 2020, vol. 8, pp. 35403–35419. DOI: 10.1109/ACCESS.2020.2974752.*
14. *Oseni A., Moustafa N., Janicke H., Liu P., Tari Z., Vasilakos A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges // Journal of ACM, 2020, vol. 37, no. 4, Article 111, 35 pages. DOI: 10.1145/1122445.1122456.*
15. *Xu H., Ma Y., Liu H.C. et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review // International Journal of Automation and Computing, 2020, vol. 17, pp. 151–178. DOI: 10.1007/s11633-019-1211-x.*

16. *Ren K., Zheng T., Qin Zh., Liu X. Adversarial Attacks and Defenses in Deep Learning // Engineering, 2020, vol. 6, pp. 346–360. DOI: 10.1016/j.eng.2019.12.012.*
17. *Zhou X., Canady R., Li Y., Koutsoukos X., Gokhale A. Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair // In: Darema, F., Blasch, E., Ravela, S., Aved, A. (eds) Dynamic Data Driven Applications Systems. DDDAS 2020. Lecture Notes in Computer Science, 2020, vol. 12312, pp 102–109. DOI: 10.1007/978-3-030-61725-7_14.*
18. *Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey // IEEE Access, vol. 9, pp. 155161–155196, 2021, DOI: 10.1109/ACCESS.2021.3127960.*
19. *Zhang H., Liu B., H Wu. Smart Grid Cyber-Physical Attack and Defense: A Review // IEEE Access, vol. 9, pp. 29641-29659, 2021, DOI: 10.1109/ACCESS.2021.3058628.*
20. *Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. A survey on adversarial attacks and defences // CAAI Transactions on Intelligence Technology, 2021, vol. 6, pp. 25–45. DOI: org/10.1049/cit2.12028.*
21. *Rosenberg I., Shabtai A., Elovici Y., Rokach L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain // ACM Computing Surveys, 2021, vol. 54, no. 5, Article 108, 36 pages. DOI: 10.1145/3453158.*
22. *Tian J., Wang B., Li J., Konstantinou C. Adversarial attack and defense methods for neural network based state estimation in smart grid // IET Renewable Power Generation, 2021, vol. 16, no. 16, pp. 3507-3518. DOI: 10.1049/rpg2.12334.*
23. *Kong Z., Xue J., Wang Y., Huang L., Niu Z., Li F., Meng W. A Survey on Adversarial Attack in the Age of Artificial Intelligence // Wireless Communications and Mobile Computing. 2021. Vol. 2021, Article ID 4907754, 22 pages. DOI: 10.1155/2021/4907754.*
24. *Zhou Sh., Liu Ch., Ye D., Zhu T., Zhou W., Yu Ph.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity // ACM Computing Surveys, 2022, vol. 55, no. 8, Article 163, 39 pages. DOI: 10.1145/3547330.*
25. *Khamaiseh S.Y., Bagagem D., Al-Alaj A., Mancino M., Alomari H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification // IEEE Access, vol. 10, pp. 102266–102291, 2022, DOI: 10.1109/ACCESS.2022.3208131.*
26. *Liang H., He E., Zhao Y., Jia Z., Li H. Adversarial Attack and Defense: A Survey // Electronics, 2022, vol. 11, 1283. DOI: 10.3390/electronics11081283.*
27. *Tian Q., Zhang S., Mao Sh., Lin Y. Adversarial attacks and defenses for digital communication signals identification // Digital Communications and Networks, 2022, in press. DOI: 10.1016/j.dcan.2022.10.010.*
28. *Anastasiou Th., Karagiorgou S., Petrou P., Papamartzivanos D., Giannetsos Th., Tsirigotaki G., Keizer J. Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems // Sensors, 2022, vol. 22, 6905. DOI: 10.3390/s22186905.*
29. *Li Y., Cheng M., Hsieh Ch. -J., Lee Th. C. M. A Review of Adversarial Attack and Defense for Classification Methods // The American Statistician, 2022, vol. 76, No. 4, pp. 329–345. DOI: 10.1080/00031305.2021.2006781.*
30. *Tian J., Wang B., Li J., Wang Z. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid // IEEE Transactions on Network Science and Engineering, 2022, vol. 9, no. 2, pp. 807–819. DOI: 10.1109/TNSE.2021.3135565.*
31. *Li H., Namiot D. A Survey of Adversarial Attacks and Defenses for Image Data on Deep Learning // International Journal of Open Information Technologies, 2022, vol. 10, no. 5, pp. 9–16.*
32. *Girdhar M., Hong J., Moore J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models // IEEE Open Journal of Vehicular Technology, 2023, vol. 4, pp. 417–437. DOI: 10.1109/OJVT.2023.3265363.*
33. *Goyal Sh., Doddapaneni S., Khapra M. M., Ravindran B. A Survey of Adversarial Defenses and Robustness in NLP // ACM Computing Surveys, 2023, vol. 55, no. 14s, Article 332, 39 pages. DOI: 10.1145/3593042.*
34. *Al-Khassawneh Y. A. A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges // Indonesian Journal of Science and Technology, 2023, vol. 8, no. 1, pp. 79–96. DOI: 10.17509/IJOST.V8I1.52709.*
35. *Sun L. et al. Adversarial Attack and Defense on Graph Data: A Survey // IEEE Transactions on Knowledge and Data Engineering, 2023, vol. 35, no. 8, pp. 7693–7711. DOI: 10.1109/TKDE.2022.3201243.*

ОБНАРУЖЕНИЕ АТАК НА ВЕБ-ПРИЛОЖЕНИЕ С ПОМОЩЬЮ САМООРГАНИЗУЮЩИХСЯ КАРТ КОХОНЕНА

Долгачев М. В.[1], Москвичев А. Д.[2], Москвичева К. С.[3]

DOI: 10.21681/2311-3456-2024-1-38-44

Цель статьи: *увеличение эффективности обнаружения атак на веб-приложения.*

Метод: *использование самоорганизующейся карты Кохонена для выявления атак на веб-приложения в режиме реального времени.*

Полученный результат: *рассмотрена самоорганизующаяся карта Кохонена как средство обнаружения аномальных данных. Проанализирована возможность интеграции самоорганизующейся карты Кохонена со средством защиты веб-приложений от атак и уязвимостей, то есть с системой класса Web Application Firewall. Реализовано программное средство, позволяющее выявлять аномалии в HTTP запросах и ответах средствами самоорганизующейся карты Кохонена, подобраны параметры для нейронной сети. Выбраны метрики, извлекающиеся из HTTP запросов для анализа нейронной сетью. Произведена интеграция реализованного программного средства с веб-сервером NGINX средствами модуля NGX JavaScript. Проведено функциональное и нагрузочное тестирование полученного комплекса средствами сканера безопасности OWASP ZAP. Полученные результаты позволили сделать вывод о том, что самоорганизующаяся карта Кохонена эффективно выявляет аномалии, однако ее необходимо использовать с шаблонными методами анализа.*

Практическая ценность: *в рамках исследования разработана методология тестирования средств защиты веб-приложений. Описаны программные средства, из которых состоит стенд для тестирования. Перечислены метрики, позволяющие объективно оценить эффективность средства защиты веб-приложения.*

Ключевые слова: *компьютерная атака, Web Application Firewall, защита информации, нейронные сети, анализ трафика, система обнаружения вторжений, Mutillidae.*

DETECTION OF ATTACKS ON WEB APPLICATION USING SELF-ORGANIZING KOHONEN MAPS

Dolgachev M. V.[4], Moskvichev A. D.[5], Moskvicheva K. S.[6]

Purpose of the article: *improving the effectiveness of web application attack detection.*

Method: *using self-organizing maps for real-time detection of attacks on web applications.*

The result: *The self-organizing map was considered as a means of detecting anomalous data. The possibility of integrating the self-organizing map with a web application firewall system for protection against attacks and vulnerabilities was analyzed. A software tool was implemented to detect anomalies in HTTP requests and responses using the self-organizing map. Parameters for the neural network were selected, and metrics were chosen for analysis by the neural network from the HTTP requests. The implemented software tool was integrated with the NGINX web server using the NGX JavaScript module. Functional and load testing of the integrated system was conducted using the OWASP ZAP security scanner. The results obtained led to the conclusion that the self-organizing map effectively detects anomalies, but it should be used in conjunction with template-based analysis methods.*

Practical value: *within the framework of the study, a methodology for testing web application security tools has been developed. The software that makes up the test bench is described. The metrics that allow you to objectively assess the effectiveness of the web application protection tool are listed.*

Keywords: *computer attack, Web Application Firewall, information security, OWASP, NGINX, Self-Organizing Maps, HTTP, neural networks.*

1 Долгачев Михаил Владимирович, кандидат технических наук, доцент, ФГБОУ ВО «Тихоокеанский государственный университет», г. Хабаровск, Россия. E-mail: 007428@pnu.edu.ru ORCID:0000-0003-1520-800X

2 Москвичев Антон Дмитриевич, аспирант, ФГБОУ ВО «Тихоокеанский государственный университет», г. Хабаровск, Россия. E-mail: anton.moskvichev.1996@yandex.ru ORCID: 0000-0001-6532-2463

3 Москвичева Ксения Сергеевна, студент, ФГБОУ ВО «Тихоокеанский государственный университет», г. Хабаровск, Россия. E-mail: 2016104073@pnu.edu.ru

4 Mihail V. Dolgachev, Ph.D. (in Tech.), Pacific National University, Khabarovsk, Russia. E mail: 007428@pnu.edu.ru. ORCID: 0000-0003-1520-800X

5 Anton D. Moskvichev, postgraduate, Pacific National University, Khabarovsk, Russia. E mail: anton.moskvichev.1996@yandex.ru ORCID: 0000-0001-6532-2463

6 Ksenia S. Moskvicheva, student, Pacific National University, Khabarovsk, Russia. E mail: 2016104073@pnu.edu.ru

Введение

Web Application Firewall (WAF) – это межсетевой экран, специально разработанный для защиты веб-приложений от различных атак и уязвимостей. Он размещается между веб-сервером и клиентами веб-приложения и контролирует входящий и исходящий трафик. Может представлять из себя как программный, так и программно-аппаратный комплекс [1].

Чаще всего WAF используют шаблонные правила, заранее предопределенные аналитиками. Основной их недостаток – отсутствие адаптивности. Правила предопределены заранее и не могут изменяться автоматически для учета новых уязвимостей или атак.

Например, если WAF использует шаблонные правила для обнаружения атак типа SQL-инъекции, они могут быть эффективными против известных шаблонов атак. Однако могут существовать новые уязвимости или варианты атак, которые не учитываются в этих шаблонах. Без обновления шаблонов правил, WAF не сможет корректно обнаруживать и, как следствие, защищать веб-приложение от новых угроз [2].

Для достижения высокого уровня безопасности рекомендуется использовать комбинацию шаблонных правил и алгоритмов машинного обучения, чтобы обеспечить более адаптивную и эффективную защиту веб-приложений.

Самоорганизующаяся карта Кохонена (или Self-Organizing Map) – это нейронная сеть без учителя, используемая для визуализации и анализа сложных данных. Самоорганизующиеся карты Кохонена применяются во многих областях и задачах анализа данных. Одна из них – обнаружение аномалий. Карта Кохонена способна выявлять аномальные или необычные образцы данных, которые не соответствуют ожидаемым моделям. Эта способность делает ее полезной для обнаружения и предотвращения аномальных событий или атак в различных областях. Интеграция карты Кохонена в качестве модуля WAF может позволить более эффективно выявлять угрозы информационной безопасности, тем самым обеспечивая более высокий уровень защиты от неизвестных ранее атак [3].

1. Самоорганизующаяся карта Кохонена, общие сведения

Самоорганизующаяся карта Кохонена представляет собой двумерную сетку, состоящую из нейронов, которые упорядочены в рамках определенной топологии. В качестве топологии могут выступать прямоугольная или шестиугольная сетки. Каждый нейрон имеет веса, которые инициализируются случайным образом и представляют собой векторы той же размерности, что и входные данные [4].

Прямоугольная сетка – это простой вариант топологии, где нейроны размещены в виде прямоугольной матрицы. В этой топологии каждый нейрон имеет четырех соседей внутри сетки. Преимущество прямоугольной сетки состоит в ее простоте и прямолинейности, что делает ее хорошим выбором для задач, где важна пространственная организация данных. Или если карта Кохонена должна быть сгенерирована с определенными ограничениями.

В шестиугольной сетке нейрон находится рядом с шестью соседями. Шестиугольная сетка обеспечивает большую симметрию и позволяет более гибко адаптироваться к структуре данных. Это особенно полезно, если данные не очень хорошо выравнены или имеют сложные взаимосвязи.

Процесс обучения нейронной сети состоит из нескольких итераций искать наилучшие соответствия между образцами данных и нейронами карты. Он проходит через следующие шаги:

1. ***Инициализация весов***. Это процесс, при котором веса нейронов инициализируются случайным образом.

При инициализации весов в самоорганизующейся карте Кохонена, обычно используется случайная инициализация. Каждый нейрон имеет вектор весов, который определяет его положение на карте. Вектор весов инициализируется случайными значениями, чтобы обеспечить случайный разброс нейронов на карте.

Типичным способом инициализации является выбор случайного значения для каждой компоненты вектора весов из некоторого диапазона. Часто используется равномерное распределение случайных чисел или гауссовское распределение. Например, значения весов могут быть случайно выбраны из интервала [0,1] или из стандартного нормального распределения.

Случайная инициализация весов позволяет нейронам начать адаптироваться к различным образцам данных в процессе обучения и постепенно организовывать их на карте. В дальнейшем веса будут изменяться и обновляться в соответствии с входными данными и принципами алгоритма обучения самоорганизующейся карты Кохонена.

2. ***Выбор образца данных***. Главная цель этого шага – выбрать случайный образец данных из обучающей выборки для дальнейшего сопоставления с нейронами на карте.

Обычно образец данных выбирается случайным образом из доступного набора данных. Это может быть случайный элемент из обучающего набора данных, выбранный равномерно или случайным образом с заданными вероятностями.

Например, если имеется набор данных в виде матрицы, то выбор образца может быть выполнен как выбор случайной строки из матрицы для текущей

итерации. Другим примером может быть выбор случайного изображения из набора изображений.

Выбор образца данных происходит для каждой итерации обучения, повторяя этот процесс, пока не пройдет заданное число итераций или не будет достигнуто условие остановки. Каждый выбранный образец данных затем используется для обновления весов нейронов на карте и позволяет картам Кохонена соответствовать структуре данных.

3. ***Выбор победителя***. Расстояние между входным образцом и весами каждого нейрона вычисляется, и выбирается нейрон с наименьшим расстоянием (наиболее близкий к входному образцу). Этот нейрон называется «победителем».

Для вычисления расстояния между входным образцом и весами каждого нейрона на самоорганизующейся карте Кохонена обычно используется евклидово расстояние. Однако в зависимости от задачи и требований могут использоваться и другие метрики расстояния. Формула для вычисления евклидового расстояния между векторами u и v длиной n выглядит следующим образом:

$$d(u,v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}, \quad (1)$$

где $u[i]$, $v[i]$ – i-е элементы векторов u и v.

Другие часто используемые метрики расстояния включают в себя:

- ✓ Манхэттенское расстояние (или L1-норма), которое вычисляется как сумма абсолютных разностей между координатами векторов:

$$d(u,v) = \sum_{i=1}^{n} |u_i - v_i|, \quad (2)$$

- ✓ Косинусное расстояние:

$$d(u,v) = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \cdot \sqrt{\sum_{i=1}^{n} v_i^2}} \quad (3)$$

Выбор метрики расстояния зависит от характеристик данных и целей задачи. Обычно для расчета расстояний используются готовые функции или методы, предоставляемые программными библиотеками для работы с векторными операциями или машинным обучением.

4. **Обновление весов**. Веса победителя и его соседей обновляются таким образом, чтобы они становились более похожими на входной образец. Это позволяет картам Кохонена адаптироваться к структуре данных и организовывать схожие образцы рядом друг с другом на карте.

Обновление весов в самоорганизующейся карте Кохонена осуществляется после выбора победителя и его ближайших соседей. Этот процесс позволяет нейронам адаптироваться к входным данным и организовывать себя на карте более эффективно.

Процесс обновления весов нейронов может быть описан следующим образом:

1) Определение радиуса окрестности: на основе текущей итерации обучения и параметров, таких как скорость обучения, определяется радиус окрестности вокруг победителя, в которой будет происходить обновление весов. Радиус окрестности обычно начинается с некоторого начального значения и уменьшается по мере продвижения обучения.
2) Обновление весов победителя и его соседей. Веса победителя и его ближайших соседей обновляются на основе формулы, которая обычно учитывает расстояние между нейронами и входным образцом данных. Обновление весов происходит таким образом, чтобы более похожие нейроны были более привлекательными для входных данных и постепенно приближались к ним.
3) Обновление весов остальных нейронов. Веса остальных нейронов на карте также обновляются в соответствии с их расстоянием от победителя. Чем ближе нейрон к победителю, тем больше он будет обновлять свои веса.

Обновление весов осуществляется в каждой итерации обучения и повторяется для каждого входного образца данных. Постепенно, на основе локальной конкуренции и сотрудничества между нейронами, самоорганизующаяся карта Кохонена эффективно организует и группирует данные на карте.

5. **Уменьшение шага обучения**. По мере продвижения обучения, шаг обновления весов уменьшается, позволяя сети сходиться к устойчивым состояниям.

Уменьшение шага обучения в самоорганизующейся карте Кохонена является важным процессом, который происходит по мере продвижения обучения. Это позволяет более медленно обновлять веса нейронов по мере приближения к стабильным состояниям карты.

Обычно уменьшение шага обучения осуществляется постепенно и зависит от номера текущей итерации обучения. Чаще всего применяются два подхода для уменьшения шага обучения:

1) Линейное уменьшение. Шаг обучения уменьшается линейно по мере продвижения итераций. Например, на каждой итерации шаг обучения может быть уменьшен на фиксированную величину или в масштабируемом показателе, предназначенном для учета скорости сходимости.
2) Экспоненциальное уменьшение. Шаг обучения экспоненциально уменьшается с увеличением числа итераций или эпох обучения. Это может быть реализовано, например, с использованием экспоненциального коэффициента снижения, который определяет скорость уменьшения шага обучения.

Уменьшение шага обучения позволяет улучшить стабильность и сходимость обучения. По мере приближения к конечному состоянию карты Кохонена, уменьшение шага обучения помогает избежать сильных флуктуаций весов и способствует более плавному обновлению. Это особенно важно при работе с большими объемами данных или сложными пространствами признаков.

После завершения обучения карта Кохонена может быть использована для кластеризации и визуализации данных. Близкие нейроны на карте обозначают похожие образцы, а удаленные нейроны представляют различные образцы.

Самоорганизующаяся карта Кохонена относится к классу нейронных сетей без учителя и часто применяется в задачах анализа данных, обработки изображений, компьютерного зрения, кластеризации и картирования. Она позволяет нам получить интуитивное понимание сложных данных и их структуры.

2. Использование нейронных сетей в системах WAF

Анализ трафика веб-приложения в WAF может быть выполнен несколькими способами:

1. Разбор и обработка заголовков. WAF анализирует заголовки HTTP-запросов и ответов, чтобы получить информацию о клиенте, сервере, используемых методах запроса, типах содержимого и дополнительных параметрах. Это позволяет WAF распознавать типичные признаки атак или потенциально опасные конфигурации.
2. Парсинг запросов. WAF разбирает содержимое HTTP-запросов, чтобы достать параметры, URL-адреса, куки и другую информацию. Это помогает определить, с помощью каких параметров пользователь обращается к веб-приложению и выявить потенциально опасные или злонамеренные значения.
3. Анализ содержимого. WAF анализирует тело HTTP-запросов и ответов, чтобы распознать злонамеренный код, вредоносные скрипты или другие опасные содержимые. Для этого может быть применен алгоритм обнаружения аномалий, сигнатурное сопоставление или даже применение машинного обучения для определения необычного поведения и атак.
4. Фильтрация данных. WAF может применять предопределенные фильтры и правила для проверки параметров запросов и данных на соответствие безопасности. Например, он может блокировать запросы, содержащие SQL-инъекции, XSS-атаки или другие уязвимости. Фильтрация может осуществляться с использованием регулярных выражений, ключевых слов или других подходов.
5. Обнаружение и предотвращение атак. С помощью методов обучения, анализа аномалий или сигнатурное соответствие, WAF может обнаруживать и предотвращать известные и неизвестные атаки на веб-приложение. Это может включать обнаружение SQL-инъекций, подбора паролей, кросс-сайтовой сценарной атаки и других видов атак.

Анализ трафика WAF происходит в реальном времени, поэтому система должна быть очень производительной и быстрой для эффективной защиты веб-приложения от угроз безопасности и атак [5].

WAF служит важным инструментом для обеспечения безопасности веб-приложений, предотвращения атак и обнаружения уязвимостей. Он работает как дополнительный слой защиты, помогающий предотвратить утечку данных, повреждения сайта и другие проблемы безопасности [6].

Карту Кохонена можно использовать в системах WAF для обнаружения и предотвращения атак на веб-сайты. Вот несколько способов [7]:

1. Обнаружение аномального трафика. Карта Кохонена может быть обучена на нормальном трафике веб-сайта, чтобы выявить типичные образцы поведения пользователей. Затем, в реальном времени, она может анализировать входящий трафик и идентифицировать аномальное поведение, которое может указывать на попытку атаки. Например, если входящий запрос отличается от типичных запросов, нейронная сеть может сигнализировать о возможном взломе или атаке на сайт.
2. Отслеживание и обнаружение атак. Карта Кохонена может быть обучена различать образцы трафика, связанные с известными атаками, такими как SQL-инъекции, XSS, CSRF и другие. Она может быть настроена на обнаружение характеристических признаков, связанных с такими атаками, и предотвращение их выполнения или блокировку соответствующего входящего трафика.
3. Защита от ботов и сканеров уязвимостей. Карта Кохонена может помочь в обнаружении ботов и сканеров уязвимостей, которые могут искать уязвимости веб-сайта для дальнейших атак. Используя способность к распознаванию образцов и аномалий в трафике, нейронная сеть может помочь в отслеживании и блокировке подозрительных активностей.
4. Анализ логов и инцидентов. Нейронная сеть может быть использована для анализа логов событий веб-приложения и обнаружения аномалий, нетипичного поведения или паттернов, связанных с безопасностью. Она может помочь в обнаружении новых уязвимостей, атак или аномальных событий, которые могут оставаться незамеченными при использовании традиционных методов анализа логов.

Важно отметить, что нейронная сеть должна быть обучена на репрезентативных данных и регулярно обновляться, чтобы адаптироваться к изменяющимся угрозам. Она может использоваться в сочетании с другими методами обнаружения и защиты, образуя комплексную систему безопасности веб-приложений [8, 9].

3. Применение нейронной сети Кохонена для выявления атак на веб-приложение

На рисунке 1 изображена схема работы WAF, построенного на микросервисной архитектуре и использующего нейронную сеть Кохонена для выявления атак [10, 11]. Клиент передает запрос на внешний веб-сервер NGINX. NGINX средствами модуля NJS направляет запрос модулю анализа SOM и целевому веб-приложению. Веб-приложение возвращает ответ веб-серверу NGINX. NGINX средствами модуля NJS направляет ответ модулю анализа SOM и клиенту. Модуль анализа SOM выполняет две функции: сообщает о результатах анализа аналитику в режиме реального времени и записывает запросы, ответы и результаты анализа в базу данных Storage [12, 13].

NJS (Ngx JavaScript) – это встроенный язык программирования, который используется в веб-сервере nginx для написания дополнительной логики обработки запросов и управления конфигурацией сервера. NJS базируется на языке JavaScript и предоставляет возможность расширения функциональности сервера nginx путем выполнения JavaScript-кода на этапе обработки запросов [14].

Первичное обучение происходит по данным, полученным от клиентов без анализа и фильтрации. Переобучение реализовано через базу данных. То есть нейронная сеть в момент инициализации подключается к базе данных, получает список запросов, помеченных как легитимные, затем обучается на них и анализирует входящий от клиентов трафик.

Описанная схема WAF не подразумевает активную защиту, то есть блокировку запросов и ответов, так как нейронные сети имеют высокую вероятность ложных срабатываний, что может повлечь негативное влияние на бизнес-процессы.

Модуль анализа SOM представляет из себя приложение, написанное на языке программирования Golang, реализующее работу нейронной сети Кохонена [15]. Каждый модуль схемы рисунка 1 – отдельное программное средство, запущенное в контейнере docker [16]. Приложение строит нейронную сеть со следующими параметрами:

- ✓ тип сетки – плоская;
- ✓ топология – шестиугольник;
- ✓ размер сетки – 20 на 20;
- ✓ инициализация весов производится значениями типа int, верхний предел – 500;
- ✓ выбор победителя осуществляется по евклидовому расстоянию;
- ✓ функция, используемая для обновления весов – Гауссова;
- ✓ уменьшение шага обучения – экспоненциальное [17].

Для проекции запросов к веб-приложению на числовой вектор были выбраны метрики, которые могут кардинально отличаться, если сравнить легитимный и вредоносный запросы. Это неполный список метрик, их можно расширить, однако в эксперименте брались именно эти:

- ✓ количество букв в поле запроса;
- ✓ количество цифр в поле запроса;
- ✓ количество специальных символов в поле запроса;
- ✓ общее количество символов;
- ✓ энтропия запроса.

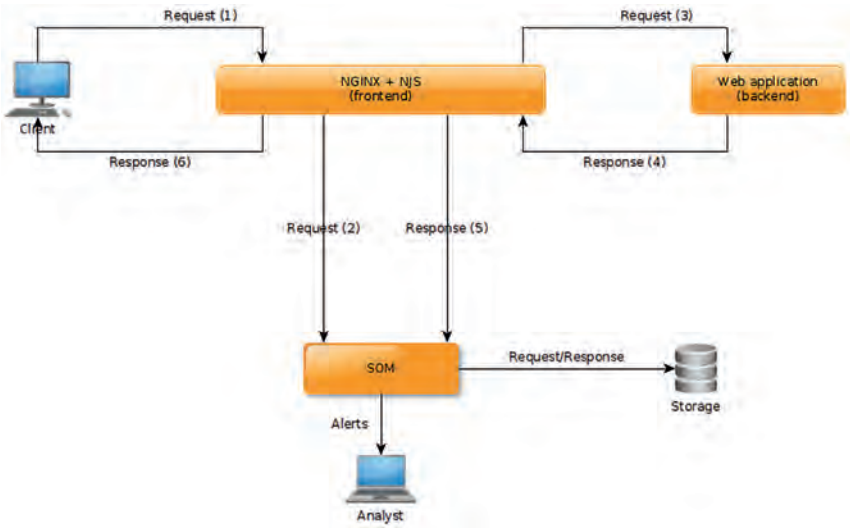


Рис. 1. Схема работы WAF, использующего нейронную сеть Кохонена

Модуль анализа SOM работает по следующему алгоритму:

1. Получает из базы или в режиме реального времени запросы клиентов к веб-приложению;
2. По запросам вычисляет необходимые для анализа метрики, формирует числовые векторы;
3. Нейронная сеть обучается на полученных векторах, в результате получаем карту Кохонена и эталонную ошибку квантования.
4. Для каждого запроса клиента в режиме реального времени вычисляется ошибка квантования. Если они выше эталонной, то запрос считается нелегитимным.

Ошибка квантования в самоорганизующейся карте Кохонена является показателем расстояния между входным образцом данных и вектором весов победившего нейрона на карте. В процессе обучения карты Кохонена каждый входной образец данных сопоставляется с ближайшим нейроном на карте, который называется победителем. Ошибка квантования измеряет расстояние между входным образцом и вектором весов победившего нейрона в пространстве признаков. Чем ниже ошибка квантования, тем более точно карта Кохонена отображает и классифицирует входные данные. Уменьшение ошибки квантования является одной из задач обучения нейронной сети и зависит от эффективности алгоритма обновления весов в процессе обучения. Ошибка квантования является метрикой, которая помогает оценить эффективность обучения нейронной сети и может использоваться для выбора наилучшей конфигурации нейронной сети или для определения условия остановки обучения. Если ошибка квантования выше эталонной, то входные данные являются нелегитимными.

Для тестирования выбрано веб-приложение Mutillidae. Mutillidae – это веб-приложение с открытым исходным кодом, разработанное OWASP (Open Web Application Security Project) в качестве целей для тестирования и практики навыков эксплуатации уязвимостей. Оно представляет собой настраиваемое и легко устанавливаемое окружение, содержащее различные типы уязвимостей, обнаруженных в реальных веб-приложениях. Mutillidae создано с целью обучить и эмулировать реальные среды уязвимых веб-приложений, чтобы разработчики, тестировщики и исследователи могли изучить и практиковать методы обнаружения и решения уязвимостей, таких как SQL-инъекции, XSS, подделка запросов между сайтами (CSRF) и другие. Mutillidae широко используется в обучающих курсах и тренировках, связанных с безопасностью приложений и тестированием уязвимостей. Это полезный ресурс для тех, кто хочет получить опыт в обнаружении и эксплуатации уязвимостей в веб-приложениях [18].

Генератором легитимных запросов выступил модуль обхода веб-приложения сканера безопасности OWASP ZAP. Генератором вредоносных запросов выступил модуль тестирования на проникновение сканера безопасности OWASP ZAP. Для простоты в качестве входных данных взяты только запросы типа GET, поле URN [19].

OWASP ZAP (Zed Attack Proxy) – это инструмент для тестирования безопасности веб-приложений с открытым исходным кодом. Он разработан и поддерживается сообществом Open Web Application Security Project (OWASP) и предоставляет возможности для обнаружения и эксплуатации уязвимостей в веб-приложениях [20].

Результаты тестирования представлены в таблице 1. Выборка запросов недостаточно большая для эксплуатации в промышленной среде, но с учетом однородности запросов выводы об эффективности подхода носят объективный характер.

Высокая нагрузка на CPU обусловлена использованием одного ядра многоядерного процессора. По факту нагрузку нельзя считать высокой. При этом оперативная память практически не нагружена. Это связано с тем, что модуль держит только в оперативной памяти только сеть Кохонена 20 на 20.

Таблица 1.

Результаты тестирования карты Кохонена

Нагрузка на CPU, %	100
Нагрузка на оперативную память, %	<1
Число запросов, на которых обучалась нейронная сеть	1450
Число проанализированных запросов (все запросы считаются вредоносными)	7331
Среднее время анализа одного запроса, микросекунды	113
Потенциальная скорость обработки запросов, запросы в секунду	8849
Процент ошибок, %	12,5

Вывод

Самоорганизующаяся карта успешно выполняет задачу анализа HTTP трафика на наличие векторов атак и попыток эксплуатации уязвимостей. Необходимые вычислительные мощности для работы анализатора оказались низкими, что является преимуществом, так как это позволяет снизить затраты на вычислительный процесс. В целом, результаты исследования указывают на эффективность и перспективность использования самоорганизующейся карты Кохонена для защиты веб-приложений.

Чтобы повысить эффективность анализа трафика и улучшить точность обнаружения атак, рекомендуется использовать нейронную сеть совместно с шаблонным анализом. Комбинируя шаблонный анализ с нейронной сетью, можно достичь более надежного обнаружения атак и уменьшить число ложных срабатываний. Такой подход значительно повышает безопасность системы и защищает веб-приложения от широкого спектра атак. Таким образом, рекомендуется использование шаблонного анализа трафика в сочетании с самоорганизующейся картой Кохонена для достижения наилучших результатов в обнаружении атак и защите веб-приложения.

Литература

1. *Clincy, V. Web Application Firewall: Network Security Models and Configuration. / V. Clincy, H. Shahriar // IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). – 2018. – P. 835-836. – DOI 10.1109/COMPSAC.2018.00144.*
2. *Коллинз, М. Защита сетей. Подход на основе анализа данных / Майкл Коллинз; пер. с англ. А. В. Добровольская. – М.: ДМК Пресс, 2020. – 308 с.: ил. – ISBN 978-5-97060-649-0.*
3. *Остроух, А. В. Системы искусственного интеллекта / А. В. Остроух, Н. Е. Суркова. – 3-е изд., стер. – Санкт-Петербург: Лань, 2023. – 228 с. – ISBN 978-5-507-46441-8.*
4. *Dogo, E. Sensed Outlier Detection for Water Monitoring Data and a Comparative Analysis of Quantization Error Using Kohonen Self-Organizing Maps. / E. Dogo, N. Nwulu, B. Twala, C. Aigbavboa // International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) – 2018. – DOI 10.1109/COMPSAC.2018.00144.*
5. *Брюхомицкий, Ю. А. Искусственные иммунные системы в информационной безопасности [Текст] / Ю. А. Брюхомицкий; Южный федеральный университет. –Таганрог: Издательство Южного федерального университета, 2019. – 147 с. – ISBN 978-5-9275-3212-4.*
6. *Чио, К. Машинное обучение и безопасность / Кларенс Чио, Дэвид Фримэн; пер. с анг. А. В. Снастина. – М.: ДМК Пресс, 2020. – 388 с.: ил. – ISBN 978-5-97060-713-8.*
7. *Arul, E. Firmware Attack Detection on Gadgets Using Kohonen's Self Organizing Feature Maps (KSOFM). / E. Arul, P. Angusamy // Third International Conference on Smart Systems and Inventive Technology (ICSSIT). – 2020. – P. 21–26. – DOI 10.1109/ICSSIT48917.2020.9214115.*
8. *Zhukovytskyy, I. Study of Combined Approach Possibilities to Detecting Network Attacks Using Artificial Intelligence Mechanisms. / I. Zhukovytskyy, V. Pakhomova, I. Tsykalo, D. Bikovska // 12th International Conference on Dependable Systems, Services and Technologies (DESSERT). – 2022. – P. 1–4. – DOI: 10.1109/DESSERT58054.2022.10018718*
9. *Belej, O. Using Hybrid Neural Networks to Detect DDOS Attacks. / O. Belej, L. Halkiv // IEEE Third International Conference on Data Stream Mining & Processing (DSMP). – 2020. – DOI: 10.1109/DSMP47368.2020.9204166.*
10. *Дэвис, К. Шаблоны проектирования для облачной среды / Корнелия Дэвис; пер. с анг. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 388 с.: ил. – ISBN 978-5-97060-807-4.*
11. *Белл, Л. Безопасность разработки в Agile-проектах: Обеспечение безопасности в конвейере непрерывной поставки / Л. Белл, М. Брантон-Сполл, Р. Смит, Д. Бэрд ; пер. с англ. А. А. Слинкин.– М.: ДМК Пресс, 2018.– 448с. :ил. – ISBN 978-5-97060-648-3.*
12. *Айвалиотис, Д. Администрирование сервера NGINX / Д. Айвалиотис .– Москва : ДМК Пресс, 2018 .– 289 с. – ISBN 978-5-97060-610-0.*
13. *Дерек де Йонге. NGINX. Книга рецептов / Дерек де Йонге; пер. с англ. Д. А. Беликова.– М.: ДМК Пресс, 2020.– 176с.: ил. – ISBN 978-5-97060-790-9.*
14. *Muzaki, R. A. Improving Security of Web-Based Application Using ModSecurity and Reverse Proxy in Web Application Firewall / R. Muzaki, B. Obrina, M. Hasditama, H. Ritchi // International Workshop on Big Data and Information Security (IWBIS). – 2020. – P. 5–16. – DOI 10.1109/IWBIS50925.2020.9255601.*
15. *Батчер, М. Go на практике [Текст] / Мэтт Батчер, Мэтт Фарина; пер. с англ. Р. Н. Рагимова; науч. ред. А. Н. Киселев. – М.: ДМК Пресс, 2017. – 374 с.: ил.– ISBN 978-5-97060-477-9 (рус.). – ISBN 978-1-63343-007-5 (анг.).*
16. *Милл, Иан. Docker на практике / Иан Милл, Эйдан Хобсон Сейерс; пер. с англ. Д.А. Беликов. – М.: ДМК Пресс, 2020.– 516 с.: ил. – ISBN 978-5-97060-772-5.*
17. *Омельяненко. Эволюционные нейросети на языке Python / Омельяненко. – Москва : ДМК Пресс, 2020.– 311 с.*
18. *Эдриан Прутяну. Как стать хакером: Сборник практических сценариев, позволяющих понять, как рассуждает злоумышленник / Эдриан Прутяну; пер. с англ. Д. А. Беликова – М.: ДМК Пресс, 2020.– 380 с.: ил. – ISBN 978-5-97060-802-9.*
19. *Shubham, L. Secure Web development using OWASP Guidelines. / L. Shubham, Kumar A., Dr. T. Subbulakshmi // 5th International Conference on Intelligent Computing and Control Systems (ICICCS). – 2021. – P. 323–332. – DOI 10.1109/ICICCS51141.2021.9432179.*
20. *Fredj, O. An OWASP Top Ten Driven Survey on Web Application Protection Methods. / O. Fredj, O. Cheikhrouhou, M. Krichen, H. Hamam, A. Derhab // Risks and Security of Internet and Systems. – 2021. – P. 235–252. – DOI 10.1007/978-3-030-68887-5_14.*

О ВЕРОЯТНОСТНОМ ПРОГНОЗИРОВАНИИ РИСКОВ В ИНФОРМАЦИОННОЙ ВОЙНЕ. Часть 2. МОДЕЛЬ, МЕТОДЫ, ПРИМЕРЫ

Манойло А. В.[1], Костогрызов А. И.[2]

DOI: 10.21681/2311-3456-2024-1-45-60

Настоящая 2-я часть работы является окончанием статьи, опубликованной в №6(58) 2023 г.

Цель 2-й части работы: *предложить модель и методы для вероятностного прогнозирования частных и интегрального рисков в информационной войне (ИВ), с их помощью на основе отдельных ретроспективных данных на примерах подтвердить работоспособность модели и методов и провести системный анализ выявленных возможностей по управлению рисками в ИВ.*

Методы исследования включают *методы теории вероятностей, методы системного анализа. В качестве моделируемой системы формально могут выступать виртуальная репутация государства, его руководства и иных представителей власти в условиях реализации разнородных угроз в ИВ. Получаемые результаты математического моделирования операций и контроопераций ИВ используются в интерпретации к исходной системе, в интересах которой проводятся соответствующие расчеты.*

Результат работы: *на основе результатов анализа стратегий операций и контроопераций (в 1-й части статьи) предложены модель и методы для вероятностного прогнозирования частных и интегрального рисков в ИВ. Результаты математического моделирования операций и контроопераций ИВ представлены на количественном уровне прогнозов в терминах вероятностей «успеха» и «неудачи» в зависимости от конкретных исходных данных, формируемых по фактам или оцениваемых гипотетически. Тем самым создана математическая основа для анализа развития информационных операций и возможных способов противодействия им на уровне получаемой в результате моделирования более адекватной функции распределения времени между соседними нарушениями системной целостности.*

В работе изучены возможности по востребованным способам противодействия операциям противника в ИВ с указанием достижимых количественных оценок и рациональных способов эффективного управления рисками. На основе их применения проанализированы примеры, иллюстрирующие работоспособность предложенного подхода.

Научная новизна: *впервые предложены количественные методы прогнозирования рисков, связанных с целенаправленными усилиями противника по дискредитацией репутации государства, его руководства и иных представителей власти в глазах мирового сообщества. Для условий неопределенности формализованы способы противодействия угрозам в квазиреальном масштабе времени. Выявлены достижимые границы в превентивном управлении рисками при ведении ИВ.*

Ключевые слова: *вероятность, репутация, прогнозирование, риск, системный анализ, угроза.*

ON PROBABILISTIC FORECASTING OF RISKS IN INFORMATION WARFARE. Part 2. MODEL, METHODS, EXAMPLES

Manoilo A. V.[3], Kostogryzov A. I.[4]

This 2nd part of the work is the end of an article published in No6_2023.

The purpose of the 2nd part *of the work is to propose a model and methods for probabilistic forecasting of particular and integral risks in information warfare (IW), with their help to confirm their usability and to conduct a systematic analysis of the identified opportunities for risk management in IW on the basis of individual retrospective data on examples.*

1 Манойло Андрей Викторович., доктор политических наук, кандидат физико-математических наук, профессор МГУ им. М. В. Ломоносова, профессор факультета политологии МГУ им. М. В. Ломоносова. E mail: Cyberhurricane@yandex.ru

2 Костогрызов Андрей Иванович, доктор технических наук, профессор, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. E mail: Akostogr@gmail.com

3 Andrey V. Manoilo, Dr. Sc. of Political Sciences, Ph. D. of Physical and Mathematical Sciences, Professor of Lomonosov Moscow State University, Professor of the Faculty of Political Science of Lomonosov Moscow State University. E-mail: Cyberhurricane@yandex.ru

4 Andrey I. Kostogryzov, Dr.Sc. of Technical Sciences, Professor, Federal Research Center «Informatics and Control» of the Russian Academy of Sciences. E-mail: Akostogr@gmail.com

Research methods include *methods of probability theory, methods of system analysis. Formally, the virtual reputation of the state, its leadership and other representatives of the authorities in the context of the implementationof heterogeneous threats in the IW can act as a simulated system. The obtained results of mathematical modeling of IW operations and counteroperations are used in the interpretation of the initial system, in the interests of which the corresponding calculations are carried out.*

Results: *based on the results of the analysis of operations and counteroperations strategies (in the 1st part of the article), a model and methods for probabilistic forecasting of particular and integral risks in IW are proposed. The results of mathematical modeling of operations and counter-operations are presented at the quantitative level of forecasts in terms of the probabilities of «success» and «failure» depending on specific initial data generated by facts or estimated hypothetically. Thus, a mathematical basis has been created for analyzing the development of information operations and possible ways to counteract them at the level of a more adequate time distribution function between neighboring violations of system integrity obtained as a result of modeling.*

The paper examines the possibilities for popular ways to counter operations in the IW, indicating achievable quantitative estimates and rational ways to effectively manage risks. Based on their application, examples illustrating the efficiency of the proposed approach are analyzed.

Scientific novelty: *for the first time, quantitative methods of forecasting risks associated with the purposeful enemy efforts to discredit the reputation of the state, its leadership and other representatives of the authorities in the eyes of the world community are proposed. For conditions of uncertainty, methods of countering threats on a quasi-real time scale have been formalized. Some achievable boundaries in preventive risk management in IW have been identified.*

Keywords: *probability, reputation, forecasting, risk, system analysis, threat.*

1. Введение

Сегодня воздействие разнородных угроз при ведении ИВ в международном публичном медиапространстве выражается в целенаправленных компрометирующих выдумках резонансного характера (лжефактах, лженамерениях), способствующих опорочиванию и дискредитации репутации государства, его руководства и иных представителей власти. Эта лицевая сторона ИВ видна всем потребителям информации, но без адекватного отделения «истины» от «лжи». Изучению этой лицевой стороны интерпретации событий посвящены многие политологические исследования. В отличии от этих исследований в настоящей работе представлена математическая основа для анализа развития информационных операций и возможных способов противодействия им на уровне получаемой в результате математического моделирования более адекватной функции распределения времени между соседними нарушениями системной целостности. При этом для условий реализации разнородных угроз в качестве моделируемой системы в работе выступают виртуальная репутация государства, его руководства и иных представителей власти.

В статье под информационной войной (ИВ) понимается особый вид гибридной войны, осуществляемый с применением информационных операций со стороны противника и мер противодействия (контроопераций) со стороны защищающейся стороны. ИВ охватывает управление психикой человека (его сознанием и подсознанием), и через это операции в ИВ направлены в итоге на дискредитацию репутации государства, его руководства и иных представителей власти в глазах мирового сообщества с последующим принуждением к подчинению неким «правилам» в интересах тех сторон, которые развязывают ИВ. Репутация государства, его руководства и иных представителей власти рассматривается как стихийно складывающийся в массовом общественном сознании образ государства, его руководства и иных представителей власти, отражающий характер ожидаемых от них действий или поведения внутри государства и на международной политической арене. По сути репутация — это некий ценный виртуальный актив, используемый для поддержания конкурентоспособности и эффективного развития государства и подлежащий особому хранению и защите, в т. ч. в условиях ИВ.

Цель настоящей работы состоит в предложении востребованных модели и методов для вероятностного прогнозирования частных и интегрального рисков в ИВ и с их помощью на основе отдельных ретроспективных данных – в проведении системного анализа выявленных возможностей по управлению рисками в ИВ.

В 1-й части статьи «Анализ стратегий операций и контроопераций для математического моделирования» проведен анализ основных стратегий ИВ, мер противодействия операциям ИВ (контроопераций), характера стратегических операций ИВ [1–10]. По результатам этого анализа разработаны общие положения математического моделирования для прогнозирования рисков и системного анализа выявленных возможностей по управлению рисками в ИВ. Развитие операций и контроопераций ИВ формализовано с использованием понятия моделируемой системы. Получаемые результаты математического моделирования операций и контроопераций ИВ

для моделируемой системы используются в интерпретации к исходной системе, в интересах которой проводятся соответствующие расчеты.

На основе результатов анализа, проведенного в 1-й части для условий разнородных неопределенностей, сделаны следующие обобщенные выводы применительно к математическому моделированию, проводимому в настоящей заключительной части работы:

- основные стратегии ИВ формально могут быть описаны в терминах случайных событий, характеризующих возникновение и развитие во времени возможных угроз реализации операций и контроопераций в ИВ;
- для случаев применения активных и пассивных мер противодействия угрозам. Возникновение и развитие угроз может быть привязано к оси времени и охарактеризовано:
  - возможной частотой возникновения конкретных угроз (несколько операций в год, по ретроспективным данным в среднем около 4–6 операций в год);
  - средним временем развития этих угроз до появления целевого негативного эффекта от реализации этих угроз (несколько месяцев, по ретроспективным данным в среднем около 3–7 месяцев);
  - средним временем условно приемлемого восстановления репутации (по ретроспективным данным в среднем от одного месяца до полугода);
- стратегические операции в ИВ формально могут быть описаны в виде сложной структуры генерального плана с обозначением целей на ближнесрочную, среднесрочную и долгосрочную перспективы во времени. Каждый из составных формализованных элементов этой структуры (реально разнесенных в пространстве и времени) связан с другими элементами логическими условиями и реализует конкретный фрагмент стратегии и набор операций ИВ для достижения интегральной цели дискредитации репутации государства, его руководства и иных представителей власти. Формально выполнение плана стратегической операции может быть описано в терминах случайных событий, характеризующих развитие во времени возможных угроз для элементов этой структуры, и связано для элементов логическими условиями «И», «ИЛИ» для достижения целей в ИВ.

Ниже предлагаются модель и методы для вероятностного прогнозирования частных и интегрального рисков, с их помощью на основе отдельных ретроспективных данных на примерах иллюстрируется работоспособность модели и методов и проводится системный анализ выявленных возможностей по управлению рисками в ИВ.

2. Вероятностная модель

За основу предлагаемого подхода к математическому моделированию принят подход, изложенный в разные годы в приложении к различным системам [11–20] и доведенный до реализации на уровне ГОСТ Р 59341-2021 «Системная инженерия. Защита информации в процессе управления информацией системы», ГОСТ Р 59991 «Системная инженерия. Системный анализ процесса управления рисками для системы».

С учетом неопределенностей расчет вероятностных показателей делается при условии или в предположении реальной или гипотетической повторяемости возможных событий и их независимости. Для математической формализации приняты следующие допущения:

- к началу периода прогноза целостность моделируемой системы полагается обеспеченной;
- для различных вариантов развития угроз существуют технологии и меры для выявления признаков возникновения источников угроз и воспрепятствования реализации угрозам (например, с использованием контроопераций), а также следов реализации угроз.

Кроме того, делается предположение о наличии возможностей по определению предпосылок к реализации угроз, а также возможностей по приемлемому восстановлению нарушаемых условий для моделируемой системы (с точки зрения противодействия операциям ИВ). Обоснованное использование выбранных мер противодействия операциям ИВ является предупреждающими контрмерами (контроперациями).

За основу формализации принят следующий поэтапный алгоритм возникновения и реализации угроз для моделируемой системы: сначала возникает источник угрозы и начинает инициироваться. Например, выполняется одно или несколько действий или вбрасывается информация, прямо или косвенно влияющие на репутацию государства или его руководства – практическим примером могут служить первые действия по «Делу об отравлении Скрипалей» и соответствующие вбросы в СМИ по стратегии «Игры с пошаговым повышением ставок» с 4 по 15 марта 2018 г., включая выступление Т. Мэй 13 марта 2018 г., когда она предъявила России ультиматум, согласно которому Россия в течение 24 часов должна «правдоподобно объясниться» по поводу инцидента в Солсбери (т. е. публично признать свою вину в отравлении С. и Ю. Скрипалей), иначе Великобритания будет рассматривать «химическую атаку в Солсбери» как акт военной агрессии[5], интервью

5 «Тереза Мэй выдвинула Москве ультиматум, согласно которому в течение 24 часов российская сторона должна правдоподобно объясниться по поводу инцидента. Срок ультиматума истек в 03:00 мск 14 марта 2018 г.». См.: Лондон официально обвинил Россию в отравлении Скрипаля. // Lenta.ru/ 2018, 13 мар. URL: https://lenta.ru/news/2018/03/14/skripal/

Ю. Скрипаль 23 мая 2018 г., вбросы по стратегии «Загонной охоты» с 5 сентября по 8 октября 2018 г., тем самым началось развитие угрозы, выражающееся в разрастающемся воздействии на массовое общественное сознание, определяющее понятие репутации. По прошествии какого-то времени, свойственного менталитету массового общественного сознания (т. е. времени, в течение которого без опровержения вброшенной информации или иных ментальных контрдействий начинает признаваться ее достоверность с соответствующим восприятием относительно репутации государства или его руководства).

Развитие угрозы осуществляется до нарушения целостности моделируемой системы, это означает реализацию возникшей угрозы (в реальности это может означать ухудшение репутации государства или его руководства до того целевого уровня, который ставился при начале соответствующих операций ИВ). Под целостностью моделируемой системы, характеризующей «успех» в ИВ, понимается такое ее состояние, которое отвечает целевому назначению модели системы. Целостность формально считается нарушенной («неудача» в результате реализации угроз за период прогноза) лишь после перехода из элементарного состояния «целостность моделируемой системы обеспечена» в элементарное состояние «целостность моделируемой системы нарушена» (т. е. на практике какая-то существенная часть субъектов, на которые осуществляется информационно-психологическое воздействие, поверит или сделает вид, что поверит в достоверность вброшенной информации. Практическим примером в «Деле об отравлении Скрипалей» нарушением целостности моделируемой системы можно считать введенные администрацией США 22 августа 2018 г. санкции в отношении России из-за приписываемой ей причастности к отравлению 4 марта 2018 г. экс-полковника ГРУ Сергея Скрипаля и его дочери Юлии в Солсбери – со ссылкой на нарушение Россией американского закона о контроле над химическим и биологическим оружием и запрете его военного применения от 1991 года). Нарушение целостности моделируемой системы характеризует состояние «неудачи» в ИВ.

Если инициировавшийся источник угрозы был выявлен до наступления элементарного состояния «целостность моделируемой системы нарушена» и приняты адекватные контрмеры, то считается, что целостность моделируемой системы не нарушена (примером такого рода мер противодействия операциям ИВ могут служить так называемые «Скрипальские чтения», перехватившие на 48 часов информационную повестку у западных (в основном, британских, американских и немецких) и российских СМИ с 3 по 4 марта 2019 г. – в первую годовщину инцидента в Солсбери). Результатом применения очередной диагностики является восстановление нарушенной целостности моделируемой системы до условно приемлемого уровня или подтверждение целостности при отсутствии ее нарушения – см. описание на рис. 1 (например, удержание политической и экономической стабильности в России после введения тысяч санкций против нее и начальных резких падений курсов рубля может рассматриваться как восстановление нарушенной целостности моделируемой системы до приемлемого уровня в условиях сложившихся реалий).

Таким образом, сформулированная модель является ничем иным, как адаптированным случаем типовой модели опасного воздействия на защищаемую систему, описанной в [11–20] и рекомендуемой ГОСТ Р 59341, ГОСТ Р 59991.

3. Базовая модель (периодический контроль состояния целостности)

Предлагаемые модель и методы позволяют оценить вероятности сохранения целостности (слева на рис. 1) и нарушения целостности моделируемой системы (справа на рис. 1) на протяжении заданного периода прогноза. Именно эта последняя вероятность с учетом негативных последствий определяется как риск нарушения целостности моделируемой системы на протяжении заданного периода прогноза. Для моделируемой системы непревышение допустимого уровня риска является следствием достаточно частого диагностирования и применения эффективных средств диагностики и восстановления приемлемой целостности при существующих ограничениях.

Для описания процессов возникновения, развития и противодействия операциям ИВ в моделируемой системе введены обозначения исходных данных моделирования:

σ – частота возникновения источников угроз;

β – среднее время развития возникшей угрозы до ее реализации в виде нарушения целостности моделируемой системы (т. е. до перехода в элементарное состояние «целостность моделируемой системы нарушена» для этого источника угроз);

$T_{меж}$ – время между окончанием предыдущей и началом очередной диагностики целостности моделируемой системы;

$T_{диаг}$ – длительность диагностики моделируемой системы (в случае неиспользования способа повышения адекватности модели по ГОСТ Р 59341-2021, приложению В 2.4 длительность диагностики $T_{диаг}$ включает в себя среднее время восстановления нарушенной целостности моделируемой системы $T_{восст}$);

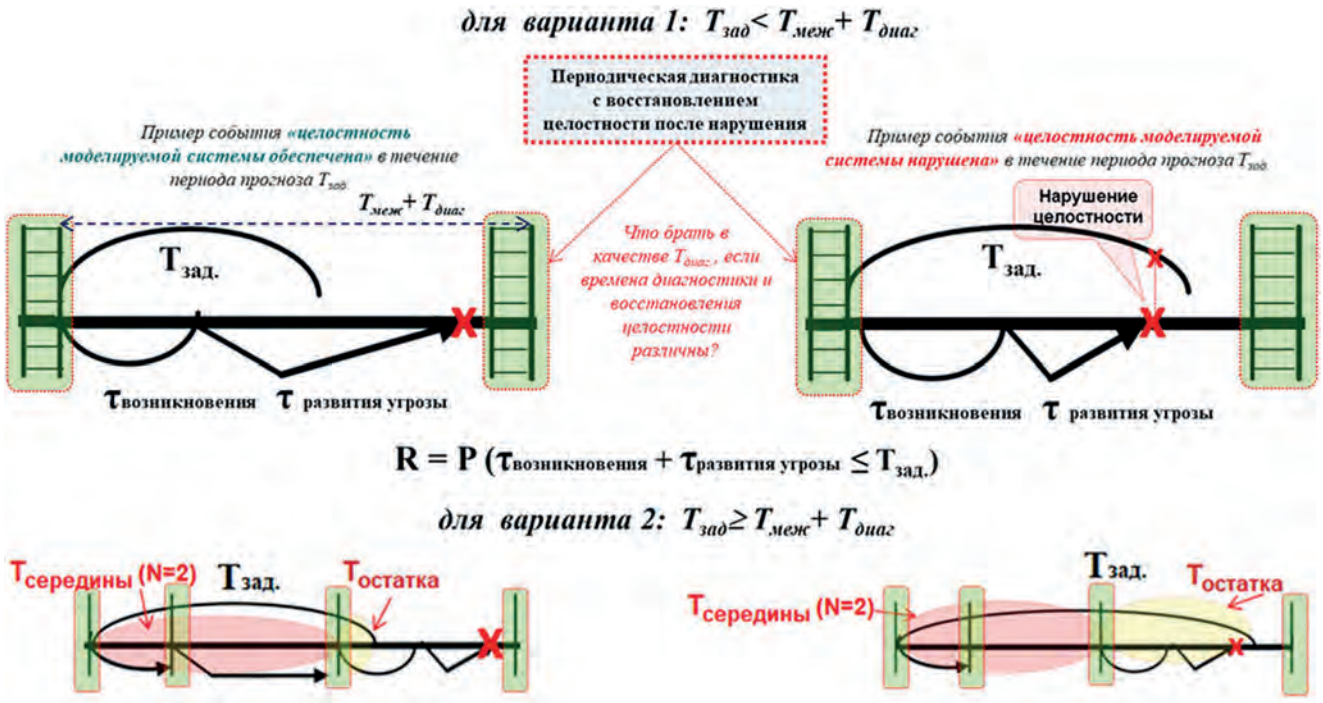


Рис. 1. Формальные случаи сохранения и нарушения целостности

$T_{восст}$ – среднее время восстановления нарушенной целостности моделируемой системы (применяется при использовании способа повышения адекватности модели по ГОСТ Р 59341-2021, приложению В 2.4);

$T_{зад}$ – длительность периода прогноза.

Для оценки вероятности нарушения целостности моделируемой системы (справа на рис. 1) $R_{наруш.}$ на протяжении заданного периода прогноза $T_{зад}$ используется выражение:

$$R_{наруш.} = 1 - P_{возд.}, \quad (1)$$

где $P_{возд.}$ – это вероятность обеспечения целостности моделируемой системы (слева на рис. 1) на протяжении заданного периода прогноза $T_{зад}$.

Возможны два варианта:

1. заданный оцениваемый период $T_{зад}$ меньше периода между окончаниями соседних диагностик ($T_{зад} < T_{меж} + T_{диаг}$);
2. заданный оцениваемый период $T_{зад}$ больше или равен периоду между окончаниями соседних диагностик ($T_{зад} \ge T_{меж} + T_{диаг}$), т.е. за это время заведомо произойдет одна или более диагностик.

Для варианта 1 вероятность $P_{возд(1)}(\sigma, \beta, T_{меж}, T_{диаг}, T_{зад})$ обеспечения целостности моделируемой системы на протяжении заданного периода прогноза $T_{зад}$ вычисляется как распределение от суммы времен возникновения и активизации опасности на момент завершения периода прогноза $T_{зад}$ – см. рис. 1:

$$P_{возд(1)} = \begin{cases} (\sigma - \beta^{-1})^{-1} \left\{ \sigma e^{-T_{зад}/\beta} - \beta^{-1} e^{-\sigma T_{зад}} \right\}, & если\ \sigma \ne \beta^{-1}, \\ e^{-\sigma T_{зад}} \left[1 + \sigma T_{зад} \right], & если\ \sigma = \beta^{-1}. \end{cases} \quad (2)$$

Эту же формулу используют для оценки вероятности обеспечения целостности моделируемой системы без какой-либо диагностики.

Для варианта 2 вероятность $P_{возд(2)}$ обеспечения целостности моделируемой системы на протяжении заданного периода прогноза $T_{зад}$. Предлагается определять по формуле (полагая, что нарушения могут произойти на срединном участке или в конце после последней диагностики до истечения длительности прогноза):

$$P_{возд(2)} = P_{серед} + P_{кон}, \quad (3)$$

где $P_{серед}$ – вероятность отсутствия нарушений целостности моделируемой системы в течение всех периодов между диагностиками, целиком вошедшими в $T_{зад}$. С учетом доли этих периодов $\frac{N(T_{меж} + T_{диаг})}{T_{зад}}$ в общем оцениваемом периоде $T_{зад}$, расчет осуществляется по формуле

$$P_{серед} = \frac{N(T_{меж} + T_{диаг})}{T_{зад}} \cdot P^{N}_{возд(1)}(\sigma, \beta, T_{меж}, T_{диаг}, T_{меж} + T_{диаг}), \quad (4)$$

N – число периодов между диагностиками, которые целиком вошли в пределы времени $T_{зад}$, $N = [T_{зад} / (T_{меж} + T_{диаг})]$ (в общем случае здесь при моделировании N – может быть действительным числом, т.е. не обязательно целым);

$P_{возд(1)}(\sigma, \beta, T_{меж}, T_{диаг}, T_{меж} + T_{диаг})$ – вероятность отсутствия нарушений целостности за один период между диагностиками, целиком вошедший в пределы времени $T_{зад}$, вычисляют по формуле (2);

$P_{кон}$ – вероятность обеспечения целостности после последней диагностики (в конце $T_{зад}$). С учетом доли

остатка $T_{ост} = T_{зад} - N (T_{меж} + T_{диаг})$ в общем периоде прогноза $T_{зад}$ расчет осуществляется по формуле

$$P_{кон} = \frac{T_{ост}}{T_{зад}} \cdot P_{возд(1)}(\sigma, \beta, T_{меж}, T_{диаг}, T_{ост}) \quad (5)$$

Значение $P_{возд(1)}(\sigma, \beta, T_{меж}, T_{диаг}, T_{ост})$ для остатка от задаваемого прогнозного периода вычисляют по формуле (2) с тем отличием, что вместо $T_{зад}$ стоит остаток $T_{ост}$.

Использование дополнительно стандартного способа повышения адекватности модели по ГОСТ Р 59341-2021, приложению В 2.4 позволяет учитывать не только среднее время системной диагностики $T_{диаг}$, но и среднее время восстановления целостности моделируемой системы $T_{восст}$.

Предложенная модель пригодна для проведения оценок системы, представимой в виде отдельного «черного ящика», причем для случая, когда времена диагностики и восстановления нарушенной целостности совпадают. Для случая, когда времена диагностики и восстановления нарушенной целостности не совпадают, предлагается использовать способ повышения адекватности, предложенный в кандидатской диссертации Нистратова А. А.[6] и доведенный до реализации в ГОСТ Р 59341, см. также рекомендации по моделированию в ГОСТ Р 59991.

Для комплексной оценки в приложении к моделируемым системам сколь угодно сложной параллельно-последовательной структуры предлагается использовать следующий алгоритм генерации новых моделей.

Рассмотрим простейшую структуру из двух независимых элементов, соединенных последовательно, что означает логическое соединение «И» (рис. 2), или параллельно, что означает логическое соединение «ИЛИ» (рис. 3). Предположение независимости имеет место быть.

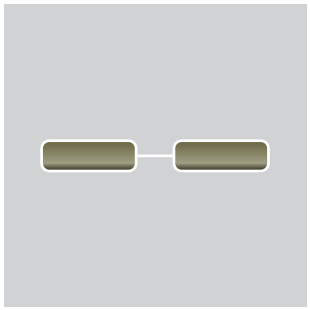


Рис. 2. Система из последовательно соединенных элементов

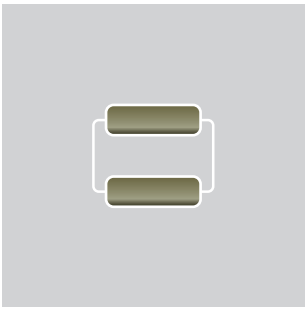


Рис. 3. Система из параллельно соединенных элементов

6 Нистратов А. А. Методика прогнозирования техногенных рисков и ее реализация с использованием Интернет-технологии. Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.17 «Теоретические основы информатики». Федеральное государственное бюджетное учреждение науки Институт проблем информатики Российской академии наук (ИПИ РАН), 2013. 150 с.

Обозначив для i-го элемента функцию распределения (ФР) времени наработки на нарушение целостности через $Bi(t) = P(\tau i \le t)$, получим:

1) для последовательно соединенных независимых элементов время до нарушения целостности равно минимуму из двух времен τi: выхода из строя 1-го или 2-го элементов (т. е. система переходит в состояние нарушенной целостности, когда откажет либо 1-й, либо 2-й элемент). В этом случае для системы в целом ФР времени наработки $B(t)$ на нарушение целостности определяется выражением

$$B(t) = P(\min(\tau 1, \tau 2) \le t) = 1 - P(\min(\tau 1, \tau 2) > t) = \\ = 1 - P(\tau 1 > t)\, P(\tau 2 > t) = 1 - [1 - B1(t)]\, [1 - B2(t)], \quad (6)$$

2) для параллельно соединенных независимых элементов (когда оба элемента находятся в функциональном состоянии и при выходе из строя одного из них другой продолжает функционировать) время до нарушения целостности равно максимуму из двух времен τi: выхода из строя 1-го и 2-го элементов, т.е. система переходит в состояние нарушенной целостности, когда выйдут из строя оба – и 1-й и 2-й элементы. В этом случае ФР времени наработки на нарушение целостности для системы в целом

$$B(t) = P(\max(\tau 1, \tau 2) \le t) = \\ = P(\tau 1 \le t)\, P(\tau 2 \le t) = B1(t)\, B2(t). \quad (7)$$

Применяя приведенные рекуррентные соотношения (6) – (7), можно получать соответствующие оценки для сколь угодно сложной логической структуры с параллельно-последовательным соединением элементов. На выходе моделирования системы – вероятность обеспечения целостности в течение заданного периода времени. Если для каждого элемента просчитать эту вероятность для всех точек $T_{зад}$. от нуля до бесконечности, то получится траектория ФР времени обеспечения целостности по каждому из элементов (или траектория, не являющаяся ФР, но близко ее аппроксимирующая) в зависимости от расчетных параметров – см. подробнее ГОСТ Р 59341-2021, приложение В.

4. Об извлечении скрытых аналитических знаний

Чтобы провести системный анализ для ответа на условный вопрос «Что будет, если...» в терминах элементарных событий за период прогноза, при формировании сценариев возможных нарушений статистика реальных событий по желанию исследователя может быть дополнена гипотетическими событиями, характеризующими ожидаемые и/или прогнозируемые условия для моделируемой системы. Применительно к анализируемому сценарию модели

ориентированы на расчет вероятности определенного элементарного состояния в течение задаваемого периода прогноза. Для негативных последствий при оценке рисков этой расчетной вероятности сопоставляют возможный материальный, моральный или репутационный ущерб. Тогда риск нарушения целостности моделируемой системы на протяжении заданного периода прогноза определен как дополнение до единицы вероятности обеспечения целостности моделируемой системы в сопоставлении с возможными последствиями относительно тех угроз, которые могут оказаться реализованными.

Согласно модели развитие критичных ситуаций в моделируемой системе считается не нарушающим ее целостности в течение заданного периода прогноза (т.е. в течение всего периода прогноза система пребывает в элементарном состоянии «целостность моделируемой системы обеспечена»), если в течение всего периода прогноза либо источники опасности не инициируются, либо после активизации происходит их оперативное выявление и принятие адекватных мер противодействия операциям ИВ. При этом, согласно допущениям, к началу прогноза моделируемая система пребывает в элементарном состоянии «целостность моделируемой системы обеспечена». Предполагается, что существуют не только средства диагностики целостности моделируемой системы, но и способы поддержания и/или ее восстановления при выявлении источников опасности или следов их активизации (см. допущения в разделе 2).

Какие скрытые знания позволяет извлечь вероятностное прогнозирование рисков?

На рис. 4 проиллюстрированы ограничения к допустимым рискам, экспоненциальная и некая более адекватная ФР времени между соседними нарушениями системной целостности с одинаковой частотой нарушений λ .

Ориентируясь на простейшую, весьма грубую, аппроксимацию экспоненциальной ФР (с одним параметром – частотой нарушений), можно легко констатировать выполнение или невыполнение задаваемых требований к уровню допустимых рисков. Ниже «пограничной полосы» – требование выполнено, выше – не выполнено. Однако это – все извлекаемые знания… Из «плюсов» – лишь удобство сравнения. И все…

Ориентируясь на более адекватную ФР или аппроксимирующую ФР функцию (например – с помощью предложенных модели и методов), если при ее создании для каждого критичного составного элемента задавались характеристики угроз и предпринимаемые меры противодействия операциям ИВ, возможно извлечение следующих знаний – см. рис. 4, 5:

- рассчитать реальную зависимость вероятности нарушения целостности системы и составных подсистем от характеристик разнородных угроз и предпринимаемых мер противодействия операциям ИВ;
- оценить точность прогнозирования по сравнению с экспоненциальной аппроксимацией ФР;
- определить период эффективного функционирования, в течение которого нарушений не ожидается (по критерию непревышения допустимых рисков) – для определения упреждающих противодействий угрозам за время, не превосходящее данного периода;
- выделить зоны прогнозных периодов времени, когда возможны нарушения требований к допустимому риску – для определения упреждающих противодействий угрозам или обоснованного управления рисками для этих зон (в т. ч. избегание рисков или смягчение требований из-за неизбежного резкого возрастания рисков в пределах, признанных приемлемыми);
- сравнить периоды эффективности, в течение которого нарушений не ожидается (по критерию непревышения допустимых рисков) с соответствующими периодами при экспоненциальной аппроксимации ФР.

Кроме этого, зафиксировав уровни «допустимых рисков» для системы и составных подсистем, а также считая неизменными все параметры, за исключением одного, возможно решение различных

Рис. 4. Фрагменты ФР, демонстрирующие экспоненциальную и более адекватную аппроксимацию

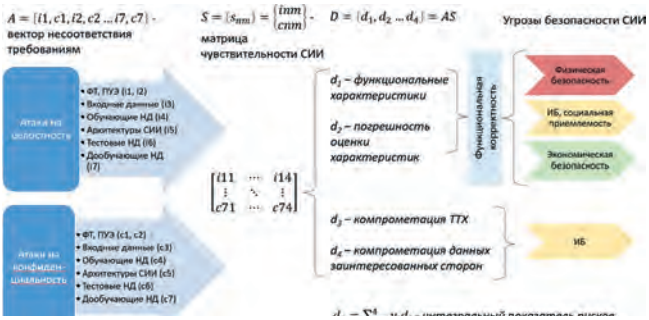


Рис. 5. Фрагменты прогнозной зависимости риска, иллюстрирующий обоснование аналитических выводов (не является ФР), см. также рис. 12

оптимизационных задач, связанных с обоснованием эффективных упреждающих мер обеспечения целостности моделируемой системы в условиях разнородных угроз.

Тем самым для условий неопределенности способы противодействия угрозам могут быть формализованы в квазиреальном масштабе времени (т.е. в масштабе времени, близком к реальному для моделируемой системы в принятых допущениях и ограничениях).

5. Примеры

Первые два примера приводятся ниже для понимания достижимого уровня рисков «фейковой» дискредитации положительной репутации виртуального политического деятеля государства в условиях, учитывающих действующее законодательство РФ. В качестве моделируемой системы в примерах выступает репутация политического деятеля. Результаты этих примеров, полученных с применением модели, аналогичной предложенной выше [18], используются далее для сравнения с другими условиями ведения ИВ по примерам 3–5.

В отличие от внутренних норм, закрепляемых в законодательствах различных государств, на международном уровне зачастую отсутствуют событийные ограничения, связанные с целями операций в ИВ и возможностями по противоборству. Цели могут оказаться долговременными, ожидаемое информационное воздействие на людей может измеряться месяцами и годами. Более того, неопределенными остаются временные характеристики развития разнородных угроз на психику человека в разных странах. В примерах 3–5 в качестве моделируемой системы выступает репутация виртуального государства и его руководства. Исходные данные сценариев ИВ сформированы с учетом ретроспективных данных в международном информационном пространстве в период с конца 2015 г. по настоящее время.

В примере 3 речь идет о виртуальном не суверенном государстве (готовом не противодействовать операциям ИВ против него или имитировать такое противодействие) и о некоем виртуальном суверенном государстве, реализующем лишь пассивные меры противодействия операциям ИВ (отрицания или оправдания, указания на нестыковки в обвинениях и т. п.), но без учета активных контрдействий. В примере 4 проводится анализ ведения ИВ виртуальным суверенным государством, реализующим не только пассивные меры противодействия операциям ИВ, но осуществляющим активные контроперации. В примере 5 проводится анализ применения длительных стратегических информационных операций в ИВ и противоборствующих контроопераций.

Пример 1 [18]. В примере осуществлен прогноз защищенности репутации виртуальных кандидатов на выборные должности от «фейков» с момента их выдвижения за 60 дней до выборов согласно законодательству РФ. Для проведения математического моделирования сформированы следующие исходные данные, учитывающие современные взгляды на характеристики «фейковых» угроз в эпоху информационно-психологического противоборства [1–5, 18] – см. табл. 1.

Результаты прогноза показали [18]: вероятностный риск дискредитации положительной репутации политика составит 0.56 в течение 1 месяца с увеличением до 0.81 в течение 2-х месяцев. Анализ показал, что сохранить изначально положительную репутацию политика в течение 2-х месяцев практически не удастся с вероятностью от 0.5 до 0.9, поскольку ожидается превалирование быстродействующих «фейков», для которых среднее время развития возникшей «фейковой» угрозы до ее реализации не будет превышать 1 месяца. При сокращении длительности судебной реакции до 2-х недель риск дискредитации изначально положительной репутации политика не будет снижаться ниже 0.6. В практической интерпретации обоснован закономерный вывод: совершенствование российского правосудия с целью сокращения до двух недель среднего времени восстановления положительной репутации добропорядочного и законопослушного политика не принесет им ожидаемой защищенности от «фейков». Вероятность дискредитации репутации политического деятеля в публичном информационном пространстве России будет соизмерима с вероятностью сохранения изначальной положительной репутации.

Таблица 1

Исходные данные для примера 1

Моделируемая система	Частота возникновения угроз, σ	Среднее время развития угроз, β	Период между диагностиками, $T_{меж}$	Длительност диагностики, $T_{диаг}$	Среднее время восстановления целостности системы, $T_{восст}$
Репутация политического деятеля	1 раз в неделю	20 суток	1 сутки	8 часов	2 недели

Таблица 2

Исходные данные для моделирования системы по примеру 2

Моделируемая система	Частота возникновения угроз, σ	Среднее время развития угроз, β	Период между диагностиками, $T_{меж}$	Длительност диагностики, $T_{диаг}$	Среднее время восстановления целостности системы, $T_{восст}$
Репутация политического деятеля	5 раз в месяц (что соизмеримо с примером 1)	20 суток (то же, что в примере 1)	1 час (вместо 1 суток для примера 1)	2 часа (вместо 8 часов для примера 1)	1 неделя (вместо 2-х недель в примере 1)

Пример 2 [18]. В примере осуществлен прогноз защищенности репутации кандидатов на выборные должности от «фейков» в период агитации за 28 дней до выборов согласно законодательству РФ. Для проведения математического моделирования сформированы следующие исходные данные – см. табл. 2.

Результаты прогноза показали: вероятностный риск дискредитации положительной репутации политика составит 0.24 в течение задаваемых 14 суток с увеличением до 0.42 в течение 28 суток (сравните с неутешительными результатами примера 1). Анализ показал, что сохранить изначально положительную репутацию политика в течение 28 суток выборной агитации сложно, но не невозможно – вероятность «успеха» может составить 0.6–0.7 против риска неудачи 0.3–0.4, т. е. вероятность «успеха» в 1.5–2 раза выше, чем риск неудачи. При сокращении сроков судебной реакции с 2-х недель до нескольких дней (от 3 до 7) риск дискредитации изначально положительной репутации политика составит в диапазоне 0.15–0.24. Эти цифры дают некоторую надежду на успешное противодействие «фейковым» угрозам.

По результатам рассмотрения примеров 1 и 2 обосновано, что наиболее эффективными на сегодня способами повышения защищенности репутации политических деятелей в РФ от «фейков» являются комплексные меры, включающие в первую очередь:

- мониторинг и выявление угроз с проведением каждый час диагностики публичного информационного пространства на предмет появления «фейков» при длительности самой диагностики не более 2-х часов;
- развитие системы правосудия и защиты репутации политического деятеля таким образом, чтобы имели место реальные возможности оперативной подачи соответствующего иска в суд при выявлении «фейка» (подача иска – за минуты) и приоритетного рассмотрения иска с тем, чтобы окончательный судейский вердикт был сформирован за несколько дней (в срок, не превышающий 7 суток) до истечения законодательных сроков агитации за политика.

На практике это достижимо с созданием и внедрением систем искусственного интеллекта, поддерживающего противодействие «фейковым» угрозам, что требует специальной научно-технической проработки. Но на международном уровне эти рекомендации не применимы. Более детально примеры по «фейковым» угрозам см. в [18].

Из результатов моделирования в примерах 1 и 2 следует, что риск на уровне 0.3 вполне может рассматриваться в качестве условно допустимого для угроз, свойственных ИВ. Справедливости ради следует отметить, что для автоматизированных систем требования к допустимым рискам гораздо более жесткие. Так, допустимая вероятность нарушения надежности предоставления информации и интегральный риск нарушения реализации процесса управления информацией системы с учетом требований по защите информации задаются на уровне 0.01–0.05, допустимая вероятность нарушения конфиденциальности информации – на уровне 0.001–0.005, а допустимая вероятность нарушения своевременности обработки запросов в системе – на уровне 0.1–0.3 (последнее – соизмеримо с полученными в примерах 1, 2 результатами прогнозов). При этом период прогноза для расчетных показателей подбирают таким образом, чтобы вероятностные значения рисков не превышали допустимые (см. ГОСТ Р 59341).

Пример 3. В примере осуществлен прогноз целостности моделируемой системы, в качестве каковой выступает репутация виртуального государства и его руководства в условиях угроз ИВ без использования каких-либо противодействующих контроопераций. В качестве аналога моделируемого сценария угроз рассмотрены, например, действия против РФ по стратегии «удушения» («Петли Анаконды») с 9 ноября 2015 г. по настоящее время с учетом пассивных мер противодействия (отрицания или оправдания, указания на нестыковки в обвинениях и т.п.), но без учета активных контрдействий со стороны РФ (каковые начали осуществляться с 3 марта 2019 г – см. 1-ю часть статьи). Учитывая разноплановость и неравномерную повторяемость разнородных угроз ИВ, исходные данные для моделирования сформированы по статистике ретроспективных данных, приведенных в 1-й части и во введении настоящей статьи.

Таблица 3

Исходные данные для примера 3

Моделируемая система	Частота возникновения угроз, σ	Среднее время развития угроз, β	Период между диагностиками, $T_{меж}$	Длительност диагностики, $T_{диаг}$	Среднее время восстановления целостности системы, $T_{восст}$
Репутация государства и его руководства	6 операций в год	3 месяца	1 сутки	8 часов	6 месяцев

Реализация угроз завершается некоторым нарушением приемлемой целостности моделируемой системы (т. е. ухудшением состояния государства, связанного с его репутацией) с полным или частичным достижением целей, которые ставились противником при начале соответствующих операций ИВ – например, до введения действительно чувствительных экономических санкций или политических воздействий. В качестве некоторых из таковых способов «удушения» в реальности были санкции, введенные в рассматриваемый период времени против РФ и ее союзников. Правдоподобные исходные данные для моделирования отражены в табл. 3.

Прежде, чем учесть все исходные данные из табл. 3, рассмотрим гипотетичный случай отсутствия какой-либо диагностики информационного пространства, пассивных и активных мер противодействия угрозам. Этот случай свойственен не суверенным государствам, готовым не противодействовать операциям ИВ против него или имитировать такое противодействие. Результаты прогноза с использованием предложенной выше модели показали: вероятностный риск нарушения целостности моделируемой системы в случае отсутствия какой-либо диагностики информационного пространства, пассивных и активных мер противодействия угрозам составит 0.69 при прогнозе на полгода с увеличением до 0.999 в течение двух лет – см. рис. 6. В практической интерпретации это означает, что в реальности первая же реализованная угроза приведет к достижению поставленных противником целей информационной операции против не суверенного государства. Поражение такого государства в ИВ неизбежно, оно будет заключаться в разрушении репутации государства и его руководителей внутри страны и на международной арене, и в полном подчинении победителю.

Суверенное государство осуществляет регулярный контроль информационного пространства и, как минимум, пассивные меры противодействия – таковыми могут быть отрицания или оправдания в информационном пространстве, указания на нестыковки в обвинениях и т. п. Кроме того, предпринимаются усилия по восстановлению нарушаемой репутации – см. рассматриваемые усредненные сценарные условия в табл. 3.

Результаты прогноза показали: вероятностный риск нарушения целостности моделируемой системы при регулярном контроле информационного пространства и пассивных мерах противодействия угрозам составит 0.26 при прогнозе на полгода с увеличением до 0.80 в течение двух лет – см. рис. 7. Пилообразность зависимости объясняется тем, что перед диагностикой с возрастанием времени риск возрастает, после диагностики – ненамного снижается с учетом возможностей восстановления после потенциальных нарушений целостности. Математически это определяется выражениями (3)–(5), а также способами повышения адекватности модели по ГОСТ Р 59341-2021, приложению В 2.4.

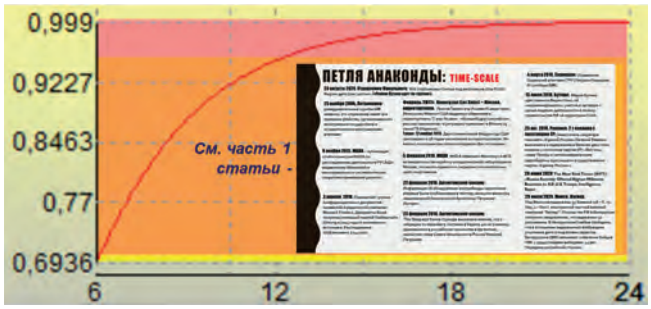


Рис. 6. Зависимость риска нарушения целостности моделируемой системы от периода прогноза (в месяцах) в случае отсутствия какого-либо контроля, пассивных и активных мер противодействия угрозам

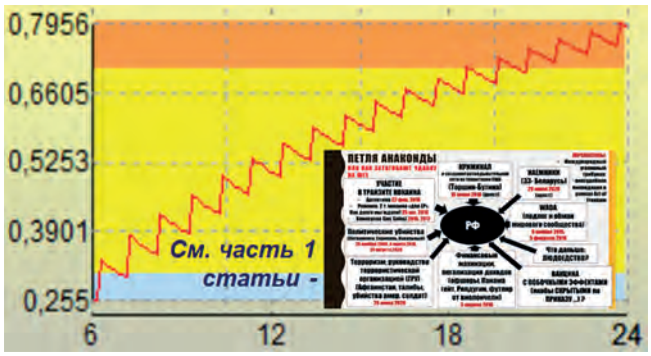


Рис. 7. Зависимость риска нарушения целостности моделируемой системы от периода прогноза (в месяцах) при регулярном контроле и пассивных мерах противодействия угрозам

В практической интерпретации приведенные на рис. 7 результаты расчетов означают, что в реальности при прогнозировании на полгода вероятность «успеха» в 3 раза больше вероятности «неудачи» (поскольку за этот сравнительно короткий срок реально может быть реализована угроза в рамках одной операции противника). При прогнозировании на год шансы «успеха» и «неудачи» оцениваются где-то «50 на 50», т.е. приблизительно равны. При прогнозировании на два года вероятность «неудачи» в 4 раза превышает вероятность «успеха» в сохранении приемлемой репутации в сценариях угроз из табл. 3. Таким образом, результаты расчетов по примеру 3 подтверждают: применение лишь пассивных мер противодействия угрозам при средне- и долгосрочных прогнозах малоперспективно, для рассмотренного сценария (см. табл. 3) в течение срока двух лет и более неизбежно последуют «неудачи» в ИВ.

Пример 4. В примере осуществлен прогноз целостности моделируемой системы, в качестве каковой выступает репутация виртуального государства и его руководства в условиях угроз ИВ с учетом пассивных и активных мер противодействия угрозам. Активные меры противодействия угрозам в ИВ заключаются в проведении контроопераций, характеристика которых приведена в 1-й части статьи.

Моделируется параллельная структура системы, представленная на рис. 3. С точки зрения защищающейся стороны формализация противодействия выглядит следующим образом: целостность моделируемой системы сохраняется, если «ИЛИ» применение пассивных мер противодействия угрозам обеспечивает «успех» (это – верхний моделируемый элемент со сценарием угроз из табл. 3 примера 3), «ИЛИ» применение активных мер противодействия угрозам с использованием контроопераций обеспечивает «успех» в ИВ (это – нижний элемент, исходные данные для него характеризуются сценарием угроз из табл. 4). На практике нарушение целостности моделируемой системы из двух логически параллельных элементов наступает только тогда, когда оба они в условиях информационных угроз, свойственных каждому элементу (а эти условия в общем случае различны), оказываются в состоянии «неудачи», т.е. в элементарном состоянии «целостность моделируемой системы нарушена» (т.е. на практике, невзирая на активные и пассивные меры противодействия какая-то существенная часть субъектов, на которые осуществляется информационно-психологическое воздействие, поверит или сделает вид, что поверит в достоверность вбрасываемой при реализации угроз информации).

Комментарии к табл. 4: в результате контроопераций согласно результатам системного анализа их ретроспективного влияния на ход ИВ (из 1-й части статьи и введения) среднее время развития угроз для нижнего элемента увеличено с 3-х до 7 месяцев, в то же время сделано предположение, что среднее время восстановления целостности снизится с полугода до 1 месяца, что является вполне правдоподобным за счет заранее предусмотренных смягчающих мер противодействия угрозам. Остальные учитываемые параметры остались неизменными по сравнению с примером 3.

Результаты прогноза показали (см. рис. 8): вероятностный риск нарушения целостности моделируемой системы при регулярном контроле информационного пространства, пассивных и активных мерах противодействия угрозам в ИВ составит 0.085 при прогнозе на полгода с увеличением до 0.45 в течение двух лет – см. рис. 8. Пилообразность зависимости объясняется так же, как и для примера 3. Несколько усиленный рост рисков при прогнозе в районе 13 и 19 месяцев по сравнению с рис. 7 объясняется тем, что среднее время развития угроз изменилось с 3-х месяцев до 7.

В практической интерпретации приведенные на рис. 8 результаты расчетов означают, что в реальности при прогнозировании на полгода вероятность «успеха» на порядок больше вероятности «неудачи»

Таблица 4

Исходные данные для нижнего элемента (см. рис. 3) по примеру 4

Моделируемый элемент системы, реализующий активные меры противодействия	Частота возникновения угроз, σ	Среднее время развития угроз, β	Период между диагностиками, $T_{меж}$	Длительност диагностики, $T_{диаг}$	Среднее время восстановления целостности системы, $T_{восст}$
Репутация государства и его руководства	6 операций в год (то же, что для верхней подсистемы)	7 месяцев (в сравнении с 3 месяцами для верхней подсистемы)	1 сутки (то же, что для верхней подсистемы)	8 часов (то же, что для верхней подсистемы)	1 месяц (в сравнении с 6 месяцами для верхней подсистемы)

Рис. 8. Зависимость риска нарушения целостности моделируемой системы от периода прогноза (в месяцах) при регулярном контроле, пассивных и активных мерах противодействия угрозам

Рис. 9. Зависимость риска нарушения целостности моделируемой системы от периода прогноза (в месяцах) для гипотетически идеального варианта

([1–0.085] / 0.085~10.8), поскольку за этот сравнительно короткий срок скорее всего не может быть реализовано ни одной угрозы, т.к. время развития угроз до их реализации в результате контроперациий увеличилось до 7 месяцев. При прогнозировании на 12–16 месяцев вероятность «успеха» в противодействии угрозам составит от 0.177 до условно допустимого уровня 0.30, что в 2.3–4.6 раза больше вероятности «неудачи». При прогнозировании на два года в сценариях угроз из таблиц 3 и 4 вероятности «успеха» и «неудачи» приблизительно одинаковы (0.55 против 0.45). Таким образом, результаты расчетов по примеру 4 показали: применение активных мер противодействия угрозам (т. е. использование контроопераций) в дополнение к пассивным мерам противодействия угрозам перспективно при кратко- и среднесрочном прогнозе ведения ИВ (до 16 месяцев). Вместе с тем, при ведении ИВ в течение двух лет и более по сценарию, приведенному в таблицах 3 и 4, «успехи» и «неудачи» приблизительно равновероятны.

При этом возникает чисто гипотетичный, но важный практический вопрос: «Какой эффективности можно добиться, если в пассивном и активном противоборстве ориентироваться на результаты, свойственные только активным мерам противодействия угрозам?», т.е. каковы могут быть самые оптимистические результаты контроопераций? На практике это означает полное информирование защищающейся стороны о планах противника (что с реальным противником не достижимо никогда). С математической точки зрения анализ этого гипотетически идеального варианта означает, что исходные данные для моделирования верхней и нижней подсистем (в структуре рис. 3) одинаковы и принимают значения из табл. 4.

Результаты прогноза для этого гипотетически идеального варианта показали (см. рис. 9): вероятностный риск нарушения целостности моделируемой системы составит 0.053 при прогнозе на полгода (что вполне сравнимо с 0.085 для вполне реального варианта из рис. 8) с увеличением до 0.41 в течение двух лет (что также вполне сравнимо с 0.45 из рис. 8). В практической интерпретации приведенные на рис. 9 результаты расчетов означают, что в идеале вероятность «неудачи» в противодействии угрозам не превысит условно допустимого уровня 0.30 при прогнозировании на срок до 20 месяцев (что вполне сравнимо с 16 месяцами из рис. 8), это как минимум в 2.3 раза меньше вероятности «неудачи» ([1 – 0.30] / 0.30 ~ 2.3).

Таким образом, результаты расчетов по примеру 4 показали: при кратко- и среднесрочном планировании даже при полном информировании о планах противника гипотетически идеальный вариант применения активных мер противодействия угрозам несущественно повышает эффективность контроопераций в ИВ. Для управления рисками правомерна рекомендация: при планировании контроопераций целесообразно ориентироваться на активные и пассивные меры противодействия угрозам с расчетом удержания эффекта от контрвоздействий в ИВ до 16 месяцев.

Пример 5. В примере осуществлен прогноз целостности сложной моделируемой системы (см. рис. 10), в качестве каковой выступает репутация виртуального государства и его руководства, подвергающихся длительным стратегическим информационным операциям ИВ и осуществляющих противоборствующие контроперации.

Рис. 10. Моделируемая система для примера 5

Подсистема 1 характеризует проведение стратегических операций в заданный период времени. Как частный случай период времени может быть равен одному году. Структура подсистемы 1 аналогична структуре, рассмотренной в примере 4. Элемент 11 ассоциируется с пассивными мерами противодействия угрозам (как в примерах 3, 4), элемент 12 ассоциируется с активными мерами противодействия угрозам (как в примере 4).

Подсистемы 2 и 3 также характеризуют проведение стратегических операций в заданный период времени. Как частный случай период времени может быть равен одному году (как и для подсистемы 1). Структура подсистем 2 и 3 аналогична структуре подсистемы 1, а также структуре, рассмотренной в примере 4. Аналогично элементы 21 и 31 также ассоциируются с пассивными мерами противодействия угрозам (как в примерах 3, 4), элементы 22 и 32 ассоциируются с активными мерами противодействия угрозам (как в примере 4). Разница между подсистемами 2 и 3 лишь в том, что за счет последствий контроопераций в период функционирования подсистемы 1 в подсистеме 2 возникает уже не 6, а 5 информационных операций со стороны противника, а в подсистеме 3 возникает уже не 6 или 5, а 4 операции со стороны противника.

Исходные данные для моделирования сформированной системы по примеру 5 отражены в табл. 5.

Важно заметить, что, единый период прогноза для всех трех подсистем при применении предложенной модели интерпретируется так: с точки зрения защищающейся стороны целостность моделируемой системы сохраняется, если за задаваемый период прогноза «И» для первой, «И» для второй, «И» для третьей подсистем их целостность не будет нарушена – то есть на практике нарушение целостности моделируемой системы из трех логически последовательных подсистем наступает тогда, когда состояние одной из них в условиях информационных угроз, свойственных каждой из подсистем и ее элементов (эти условия в общем случае различны), оказалась в элементарном состоянии «целостность подсистемы нарушена». В свою очередь, это состояние будет тогда, когда в период прогноза одновременно оба элемента окажутся в элементарном состоянии «целостность элемента нарушена» (что на практике означает достижение целей операции противником).

Это вовсе не означает, что подсистемы физически действуют в одно и то же время. Логически это могут быть разные по содержанию, но единые по длительности периоды, объединенные единым комплексом стратегических операций в ИВ – например, подсистема 1 ассоциируется с 2018 годом из «Дела Скрипалей», подсистема 2 – с 2019 годом, а подсистема 3 – с 2020 годом (см. в 1-й части статьи рис. 8). Тогда с помощью предложенной модели в примере может быть оценена возможность противодействия трехгодичному комплексу стратегических операции ИВ со стороны противника.

Результаты прогноза для защищающейся стороны показали (см. рис. 11, 12):

Таблица 5

Исходные данные для моделирования сложной системы по примеру 5

Моделируемая подсистема	Частота возникновения угроз, σ	Среднее время развития угроз, β	Период между диагностиками, $T_{меж}$	Длительност диагностики, $T_{диаг}$	Среднее время восстановления целостности системы, $T_{восст}$
Подсистема 1, состоящая из элементов 11 / 12	6/6 операций в год (как в таблице 3)	3/7 месяцев (как в таблицах 3 и 4)	1/1 сутки (как в таблицах 3 и 4)	8/8 часов (как в таблицах 3 и 4)	6/1 месяц (как в таблицах 3 и 4)
Подсистема 2, состоящая из элементов 21 / 22	5/5 операций в год (меньше, чем в подсистеме 1 за счет последствий контроопераций)	3/7 месяцев (как в таблицах 3 и 4)	1/1 сутки (как в таблицах 3 и 4)	8/8 часов (как в таблицах 3 и 4)	6/1 месяц (как в таблицах 3 и 4)
Подсистема 3, состоящая из элементов 31 / 32	4/4 операций в год (меньше, чем в подсистеме 2 за счет последствий контроопераций)	3/7 месяцев (как в таблицах 3 и 4)	1/1 сутки (как в таблицах 3 и 4)	8/8 часов (как в таблицах 3 и 4)	6/1 месяц (как в таблицах 3 и 4)

- вероятностный риск нарушения целостности подсистемы 1 за год (например, за 2018 год из «Дела Скрипалей» при 6 операциях в год) составит 0.19, подсистемы 2 – 0.15 (например, за 2019 год при 5 операциях в год), подсистемы 3 – 0.12 (например, за 2020 год при 4 операциях в год), а моделируемой системы в целом (например, за все 3 года) – 0.39. Если для каждого отдельного года результаты могут быть расценены как успешные (риски от 0.12 до 0.19 в год), то для комплекса стратегических операций за 3 года результат неутешительный (риск = 0.39), для управления рисками необходимо изыскивать политические, экономические, социальные и иные контрмеры по уменьшению частоты возникновения угроз (желательно не более 4-х операций в год), увеличению времени развития угроз до их реализации (желательно не менее 7 месяцев), а также по дальнейшему снижению времени восстановления целостности системы (в среднем для восстановления репутации за время не более 1 месяца);
- если при планировании контроопераций рассматривать периоды прогноза по полгода для каждой из подсистем, то риск нарушения целостности всей моделируемой системы составит 0.192, риск нарушения целостности моделируемой системы не превысит условно допустимого уровня 0.30, если период прогноза не превысит 9.8 месяцев. Это внушает умеренный оптимизм и уверенность в рациональности планируемых пассивных и активных мер противодействия угрозам, проявляемых в течение полугода (со значениями параметров из табл. 5). Из этих результатов для управления рисками вытекает рекомендация: при планировании контроопераций против комплекса стратегических операций противника в ИВ целесообразно ориентироваться на контроперации, применимые в течение периода до 9 месяцев, а также на иные меры противодействия угрозам с расчетом удержания эффекта от контрвоздействия в ИВ до 16 месяцев (последнее – с учетом результатов исследований в примере 4);
- если рассматривать периоды прогноза от 10 до 19 месяцев для каждой из подсистем, то риск нарушения целостности моделируемой системы составит от 0.3 до 0.7, это может быть интерпретировано как существенная неопределенность в шансах на «успех» или «неудачу» (при этом сумма всех трех периодов комплекса стратегических операций составит от 30 до 57 месяцев). Если рассматривать периоды прогноза свыше 19 месяцев для каждой из подсистем, то риск нарушения целостности моделируемой системы составит уже свыше 0.7, что может быть интерпретировано как однозначная «неудача» для защищающейся стороны. Для управления рисками правомерна рекомендация: в условиях неопределенности не планировать контроперации, ощутимый эффект от которых ожидается после 19 месяцев после их реализации (конечно, возможны исключения, связанные с заведомо определенными для государства и мира датами – такими, как годовщина важного памятного события, Олимпийские игры и т. п.).

Рис. 11. Риски нарушения целостности за год для подсистем 1–3 и моделируемой системы в целом

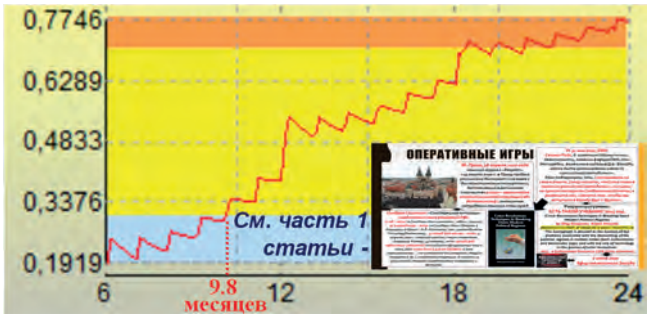


Рис. 12. Зависимость риска нарушения целостности моделируемой системы от периода прогноза (в месяцах)

Таким образом, приведенные примеры 1–5 демонстрируют работоспособность предложенных методов математического моделирования для вероятностного прогнозирования рисков. Полученные результаты моделирования и сформулированные рекомендации представляют собой аналитическую аргументацию для количественного обоснования упреждающего управления рисками в ИВ.

Заключение

1. Для математического моделирования различных способов ведения ИВ проведен анализ основных стратегий операций «удушения», «загонной охоты», прямого шантажа. В интересах аналитических исследований рассмотрены такие меры противодействия информационным операциям (меры контроопераций), как перехват информационной повестки и оперативной инициативы, отвлечение

на негодный объект, информационные прививки и контроперации возвратного типа. Проведен анализ характера стратегических операций в современной ИВ.

2. Предложены модель и методы для вероятностного прогнозирования частных и интегрального рисков в ИВ, связанных с дискредитацией репутации государства, его руководства и иных представителей власти. Для условий неопределенности способы противодействия угрозам формализованы в квазиреальном масштабе времени. Основные стратегии ИВ формально описаны в терминах случайных событий, характеризующих возникновение и развитие во времени возможных угроз реализации операций и контроперациий в ИВ. Стратегические операции в ИВ формализованы в виде сложной структуры с привязкой к генеральному плану и обозначением целей на ближнесрочную, среднесрочную и долгосрочную перспективы во времени. Каждый из составных формализованных элементов этой структуры связан с другими элементами логическими условиями и реализует конкретный фрагмент стратегии и набор операций ИВ для достижения интегральной цели противника и контроперации как меры противодействия защищающейся стороны.
3. В качестве исходных данных для математического моделирования операций и контроперациий ИВ выступают: частота возникновения источников угроз; среднее время развития возникшей угрозы до ее реализации; время между окончанием предыдущей и началом очередной диагностики целостности моделируемой системы; длительность диагностики; среднее время восстановления нарушенной целостности; длительность периода прогноза. Получаемые результаты моделирования используются в интерпретации к исходной системе, в интересах которой проводятся соответствующие расчеты.
4. Работоспособность предложенных модели и методов подтверждена на разобранных примерах, охватывающих:
   - «фейковые» воздействия на положительную репутацию виртуального политического деятеля государства в условиях, учитывающих действующее законодательство РФ, а также меры противодействия «фейкам» для удержания рисков в допустимых пределах;
   - реализацию информационных операций против виртуального несуверенного государства (готового не противодействовать операциям ИВ против него или имитировать такое противодействие) и некоего виртуального суверенного государства, реализующего лишь пассивные меры противодействия операциям ИВ (отрицания или оправдания, указания на нестыковки в обвинениях и т. п.), но без учета активных контрдействий;
   - реализацию информационных операций против виртуального суверенного государства, применяющего не только пассивные меры противодействия операциям ИВ, но и осуществляющего активные контроперации;
   - реализацию длительных стратегических информационных операций в ИВ и противоборствующих контроперациий.

   На изученных практических примерах определены достижимые границы рисков, которые могут быть использованы в поиске эффективных контрмер при ведении ИВ.
5. Математическое моделирование на примерах ведения ИВ и проведенный системный анализ зависимостей прогнозируемых рисков от исходных данных позволили выявить скрытые возможности по управлению рисками в ИВ и обосновать практические меры по удержанию рисков в допустимых пределах.

Литература

.1. *Манойло А. В., Костогрызов А. И. О вероятностном прогнозировании рисков в информационной войне. Часть 1. Анализ стратегий операций и контроперациий для математического моделирования // Вопросы кибербезопасности. 2023, №6. С. 2–19. DOI: 10.21681/2311-3456-2023-6-2-19*
2. *Манойло А. В. Фейковые новости как угроза национальной безопасности и инструмент информационного управления // Вестник Московского университета. Серия 12: Политические науки. — 2019. — № 2. — С. 41–42.*
3. *Трубецкой А. Ю. Психология репутации. – М.: Наука, 2005. – 291 с.*
4. *Устинова Н. В. Политическая репутация: сущность, особенности, технологии формирования: дис. канд. полит. наук. – Екатеринбург: УГУ, 2005. – 166 с.*
5. *Шишканова А. Ю. Репутация политического лидера: особенности и технологии формирования // Огарёв-Online. 2016. №7(72). С. 2.*
6. *Манойло А. В., Петренко А. И., Фролов Д. Б. Государственная информационная политика в условиях информационно-психологической войны. 4-е изд., перераб. и доп. — Горячая линия-Телеком Москва, 2020. — 636 с.*
7. *Манойло А. В. Современная практика информационных войн и психологических операций. Вирусные технологии и «эпидемии» каскадного типа на примере операции по разоблачению агента влияния ЦРУ, бывшего вице-президента Венесуэлы Диосдадо Кабельо 17-21/08/2019. // Национална сигурност (Nacionalna sigurnost). 2019. Выпуск №3. С. 3–8. URL: https://nacionalna-sigurnost.bg/broi-3/*

8. *Манойло А. В. Дело Скрипалей как операция информационной войны // Вестник Московского государственного областного университета. — 2019. — № 1.*
9. *Манойло А. В. Цепные реакции каскадного типа в современных технологиях вирусного распространения фейковых новостей // Вестник Московского государственного областного университета (Электронный журнал). — 2020. — № 3.*
10. *Климов С. М. Модели анализа и оценки угроз информационно-психологических воздействий с элементами искусственного интеллекта. / Сборник докладов и выступлений научно-деловой программы Международного военно-технического форума «Армия-2018». 2018. С. 273–277.*
11. *Костогрызов А. И., Степанов П. В. Инновационное управление качеством и рисками в жизненном цикле систем – М.: Изд. «Вооружение, политика, конверсия», 2008. – 404с.*
12. *Andrey Kostogryzov, Andrey Nistratov, George Nistratov Some Applicable Methods to Analyze and Optimize System Processes in Quality Management // InTech. 2012. P. 127–196. URL = http://www.intechopen.com/books/total-quality-management-and-six-sigma/some-applicable-methods-to-analyze-and-optimize-system-processes-in-quality-management*
13. *Grigoriev L., Kostogryzov A., Krylov V., Nistratov A., Nistra- tov G. Prediction and optimization of system quality and risks on the base of modelling processes // American Journal of Operation Researches. Special Issue. 2013. V. 1. P. 217–244. http://www.scirp.org/journal/ajor/*
14. *Andrey Kostogryzov, Pavel Stepanov, Andrey Nistratov, George Nistratov, Oleg Atakishchev and Vladimir Kiselev Risks Prediction and Processes Optimization for Complex Systems on the Base of Probabilistic Modeling // Proceedings of the 2016 International Conference on Applied Mathematics, Simulation and Modelling (AMSM2016), May 28-29, 2016, Beijing, China, pp. 186–192. www.dropbox.com/s/a4zw1yds8f4ecc5/AMSM2016%20Full%20Proceedings.pdf?dl=0*
15. *Костогрызов А. И. Прогнозирование рисков по данным мониторинга для систем искусственного интеллекта / БИТ. Сборник трудов Десятой международной научно-технической конференции – М.: МГТУ им. Н.Э. Баумана, 2019, сс. 220–229*
16. *Kostogryzov A., Nistratov A., Nistratov G. (2020) Analytical Risks Prediction. Rationale of System Preventive Measures for Solving Quality and Safety Problems. In: Sukhomlin V., Zubareva E. (eds) Modern Information Technology and IT Education. SITITO 2018. Communications in Computer and Information Science, vol 1201. Springer, pp.352–364. https://www.springer.com/gp/book/9783030468941*
17. *Kostogryzov A, Nistratov A. Probabilistic methods of risk predictions and their pragmatic applications in life cycle of complex systems. In «Safety and Reliability of Systems and Processes», Gdynia Maritime University, 2020. pp. 153–174. DOI: 10.26408/srsp-2020*
18. *Костогрызов А. И. Подход к вероятностному прогнозированию защищенности репутации политических деятелей от «фейковых» угроз в публичном информационном пространстве // Вопросы кибербезопасности. 2023, №3. С. 114–133. DOI:1021681/2311-3456-2023-3-114-133*
19. *Kostogryzov A., Makhutov N., Nistratov A., Reznikov G. Probabilistic predictive modeling for complex system risk assessments (Вероятностное упреждающее моделирование для оценок рисков в сложных системах). Time Series Analysis – New Insights. IntechOpen, 2023, pp. 73–105. http://mts.intechopen.com/articles/show/title/probabilistic-predictive-modelling-for-complex-system-risk-assessments*
20. *Костогрызов А. И.. Нистратов А. А. Анализ угроз злоумышленной модификации модели машинного обучения для систем с искусственным интеллектом // Вопросы кибербезопасности. 2023, №5. С.*

КОНЦЕПЦИЯ ГЕНЕТИЧЕСКОЙ ДЕЭВОЛЮЦИИ ПРЕДСТАВЛЕНИЙ ПРОГРАММЫ. Часть 1

Израилов К. Е.[1]

DOI: 10.21681/2311-3456-2024-1-61-66

Цель исследования: *развитие направления реверс-инжиниринга программ, заключающегося в преобразовании их представлений в одно из предыдущих.*

Методы исследования: *системный анализ, мысленный эксперимент, аналитическое моделирование, многокритериальная оптимизация.*

Полученные результаты: *предложена концепция генетической деэволюции представлений программы, предлагающая процесс их восстановления не обратным способом, т.е. от текущего к предыдущему, а прямым – работая с псевдо-предыдущим представлением и оценивая его близость к исследуемому текущему; принцип концепции основан на решении оптимизационной задачи с помощью генетических алгоритмов.*

В первой части статьи введена онтологическая модель предметной области, в терминах которой предложена высокоуровневая схема (де)эволюции представлений, отражающая преобразования между ними, а также внесение и обнаружение уязвимостей; дано формализованное описание процессов на схеме.

Научная новизна *заключается в качественно новой точке зрения на восстановление представлений – с помощью процесса итеративного подбора предыдущего для соответствия (после эволюции) текущему, при этом, основанного на принципах генетических алгоритмов, а также имеющего полностью формализованный вид.*

Ключевые слова: *концепция, эволюция, реверс-инжиниринг, реинжиниринг, обратная разработка, обратный инжиниринг, генетический алгоритм, уязвимость.*

THE GENETIC DE-EVOLUTION CONCEPT OF PROGRAM REPRESENTATIONS. Part 1

Izrailov K. E.[2]

The goal of the investigation: *development of the programs reverse engineering direction, which consists in transforming their state into one of the previous.*

Research methods: *system analysis, mental experiment, analytical modeling, multicriteria optimization.*

Result: *the genetic de-evolution concept of program representations has been proposed, suggesting a process for their restoration in a backway, i.e. from the current to the previous one, and direct way – working with the pseudo-previous representation and assessing its proximity to the current one being studied; the concept principle is based on solving an optimization problem using genetic algorithms.*

In the first part of the article, a subject area ontological model is introduced, in terms of which a high-level scheme for the (de)evolution of representations is proposed, reflecting transformations between them, as well as the introduction and detection of vulnerabilities; a processes formalized description in the diagram is given.

The scientific novelty *consists in a qualitatively new point of view on the representations restoration – using the iterative process selection of the previous one to correspond (after evolution) to the current one, at the same time, based on the principles of genetic algorithms, and also having a completely formalized form.*

Keywords: *concept, evolution, reverse engineering, reengineering, backward engineering, genetic algorithm, vulnerability.*

1 Израилов Константин Евгеньевич, кандидат технических наук, доцент, старший научный сотрудник лаборатории проблем компьютерной безопасности Санкт-Петербургского Федерального исследовательского центра Российской академии наук, Санкт-Петербург, ORCHID: http://orcid.org/0000-0002-9412-5693. Scopus Author ID: 56122749800. E mail: konstantin.izrailov@mail.ru

2 Konstantin E. Izrailov, Ph.D., assistant Professor, Senior Researcher of Laboratory of Computer Security Problems of St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint-Petersburg. ORCHID: http://orcid.org/0000-0002-9412-5693. Scopus Author ID: 56123238800. E mail: konstantin.izrailov@mail.ru

Введение

Реверс-инжиниринг (сокр. реинжиниринг, далее – РИ) программного обеспечения является актуальнейшим направлением в области информационных технологий [1]. Назначением РИ в частном смысле является получение исходного кода программы (а точнее псевдоисходного, лишь близкого к изначальному), как правило имеющего текстовый вид и подходящего для ручного анализа экспертом, из ее машинного кода, имеющего бинарный вид и выполняемого на ЦПУ [2]; как правило, такой процесс называется *декомпиляцией* [3, 4]. В общем смысле, понимаемом автором, РИ предназначен для восстановления метаинформации, утерянной при инжиниринге программы согласно ее жизненному циклу; например, помимо получения исходного кода из машинного возможно восстановить алгоритмы функций программы, ее архитектуру, концептуальную модель и даже (в пределе) саму идею [5, 6].

Основными целями применения РИ, естественно, в условиях отсутствия доступного исходного кода, являются следующие. Во-первых, использование программ, созданных недоверенным производителем (недружественными странами, нелегальными и криминальными организациями, недобросовестными конкурентами), приводит ко множеству информационных рисков [7], поскольку такие программы могут содержать уязвимости, делая тем самым свое функционирование отличным от заявленного и/или требуемого [8]. Для чего необходимо получение информации о фактическом функционале программы, что является побочным результатом РИ [9]. Во-вторых, для осуществления эффективного (т.е. с сохранением функционала, в кратчайшие сроки и при адекватных затратах ресурсов) импортозамещения зарубежного программного обеспечения отечественным [10, 11], что особо актуально в условиях неослабевающего санкционного давления, требуется восстановление деталей реализации имеющегося – это является основным назначением РИ. И, в-третьих, при полностью легальной разработке программ, плотно взаимодействующих с внешними библиотеками или программными продуктами, требуется детальное понимание специфики такого информационного обмена; например, форматов входных данных, кодов результата выполнения и т.п. Если разработчик сторонних продуктов обеспечивает своевременную и квалифицированную обратную связь, а также имеет «добротную» документацию, то реализация взаимодействия не представляет особой сложности. Однако в ином случае, например, когда разработчик прекратил поддержку или не выполняет своих обязательств, РИ позволит самостоятельно составить спецификацию на интерфейсы взаимодействия внешних продуктов, что даст возможность реализовать и собственные [12].

На данный момент, РИ в частном смысле (т.е. как декомпиляция машинного кода в исходный) является отдельно стоящей проблемой, не имеющей удовлетворительного научно-практического решения. И хотя существуют средства декомпиляции (наиболее популярным из которых является IDA Pro с плагином Hex-Rays [13–15]), все они поддерживают ограниченный набор ЦПУ машинного кода и выдают далеко не всегда корректный псевдоисходный код; по крайне мере, рекомендуется проверять результат их работы вручную, а также применять дополнительные автоматические средства повышения человеко-ориентированности (например, путем добавления комментариев [16]). РИ в общем же смысле в принципе оставлено без существенного внимания, поскольку получение более высокоуровневых представлений программы (например, алгоритмов) из машинного кода считается второстепенной или несущественной задачей (с чем автор категорически не согласен). При этом, понимание архитектуры программы и ее концептуальной модели (а также и самой идеи) существенно упростило бы достижение трех вышеуказанных целей применения РИ. Сам РИ является достаточно технически сложным процессом – в случае автоматизации резко снижается качество результатов и появляются ограничения в применении, а при его ручном проведении – резко возрастает время и затрачиваемые ресурсы. Одним из путей качественного повышения эффективности данного процесса, с точки зрения автора, является применение искусственного интеллекта, и, в частности, генетических алгоритмов. Для этого все преобразования программы в процесс ее прямого инжиниринга рассматриваются, как части единой эволюции программы – от представления первоначальной идеи до выполняемого машинного кода (через концептуальную модель, архитектуру, алгоритмы, исходный код); обратные же преобразования представлений, логично, могут быть названы деэволюцией программы. Далее будет описана общая концепция такой генетической деэволюции представлений программы, лежащая в основе всего авторского направления исследований.

Онтологическая модель

Для использования единой терминологии введем и опишем следующую онтологическую модель предметной области, содержащую сущности и их взаимосвязи, и представленную на рис. 1; используются следующие обозначения: прямоугольник – сущность, стрелка – взаимосвязь, зеленый фон – объект, синий фон – действие, красный фон – объект из области информационной безопасности.

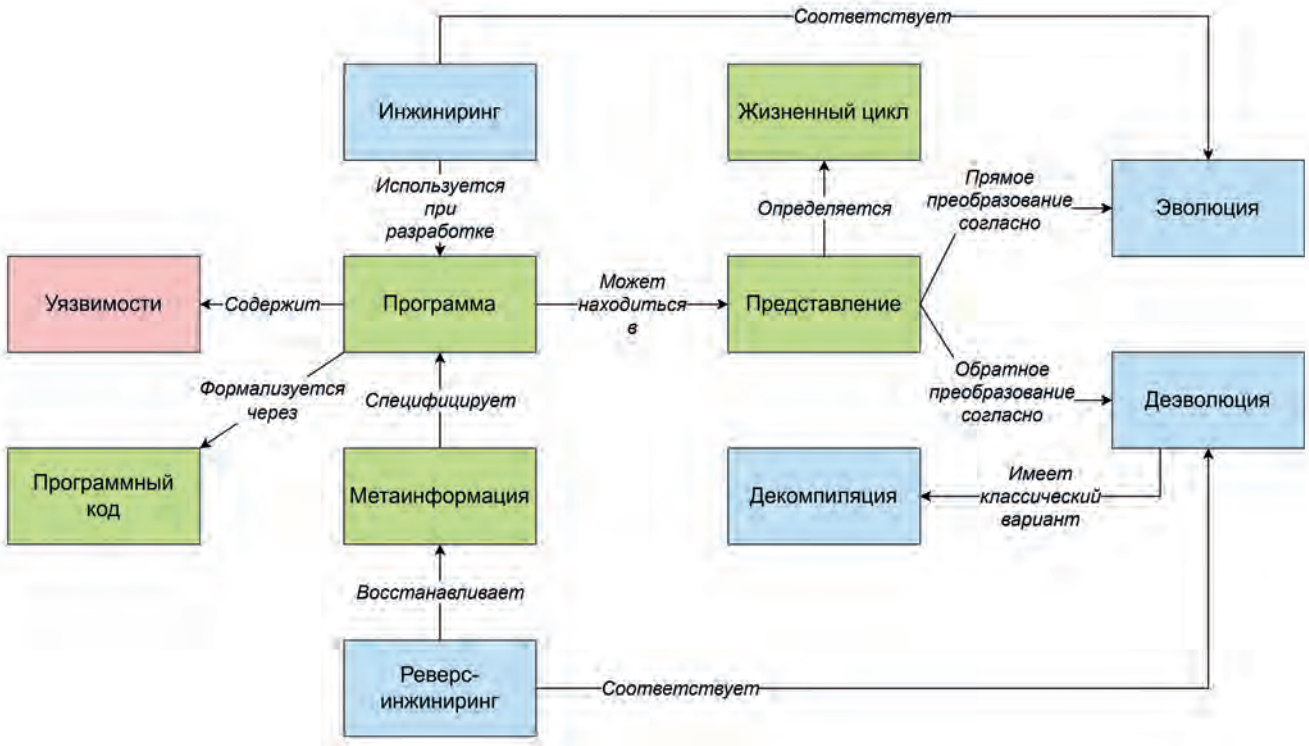


Рис. 1. Онтологическая модель предметной области

Онтологическая модель (см. рис. 1) состоит из следующих элементов (курсивом отмечены связи):

1. Программа – информационный объект, предназначенный (как в данной момент, так и в будущем) для выполнения на ЦПУ и *формализуемый через* программный код;
2. Программный код – формальная запись (текстовая, графическая, иная) реализации функционала по решению определенных задач;
3. Метаинформация о программе – информация, *специфицирующая* программу в человеко-ориентированном виде и не предназначенная напрямую для ее выполнения;
4. Представление программы – типовое состояние (как совокупность формы и содержания [17]), в котором *может находиться* программа в определенный момент своего существования;
5. Жизненный цикл программы – схема, *определяющая* закономерность смены представлений программы;
6. Эволюция представлений программы – *прямое преобразование* представлений программы согласно жизненного цикла (т.е. от первоначальной идеи к конкретной реализации);
7. Деэволюция представлений программы – *обратное преобразование* представлений программы согласно жизненного цикла (т.е. от конкретной реализации к первоначальной идее);
8. Инжиниринг программы – процесс, *используемый при разработке* программы, и соответствующий эволюции ее представлений;
9. Реверс-инжиниринг (реинжиниринг) представления программы – процесс, *восстанавливающий* метаинформацию из программы и *соответствующий* деэволюции ее представлений;
10. Декомпиляция программы – *классический вариант* деэволюции представлений программы по переходу из машинного кода в псевдоисходный код;
11. Уязвимость программы – дефект, содержащийся в программе, в виде отличия ее содержания от «идеального» (изначально задуманного).
12. Опишем предложенную онтологию в практическом аспекте. Центральным звеном модели (см. рис. 1) является программа, формализуемая через программный код и которая в процессе своего существования может находиться в некоторых представлениях (возникших исторически и/или обусловленных типовым процессом программного инжиниринга) – машинного кода, идентичного ему ассемблерного кода, исходного кода, алгоритмов, архитектуры, концептуальной модели и идеи. Каждое представление программы определяется совокупностью формы и содержания, смысл которых неоднократно объяснялся в авторских статьях [5, 6, 17]: форма – внешний вид программы, заданный с использованием определенной нотации (языка программирования, блок-схем, текстового описания и т.п.); содержание – внутренняя логика программы, определяющая ее функционал и не зависящая от способа описания (т.е. от формы). Схема

перехода между представлениями определяется жизненным циклом программы: например, в процессе компиляции исходного кода получается ассемблерный код, который после ассемблирования (и линковки) преобразуется в машинный. При инжиниринге программы, когда из ее начальной идеи в голове человека получается машинный код, готовый для выполнения на ЦПУ, происходит постепенная эволюция представлений. Обратным процессом является РИ, при котором из текущего представления программы (как правило, машинного или ассемблерного кода) получается псевдоисходный код, за которым возможно получить алгоритмы, архитектуру и т.д.; для этого, в частности, восстанавливается метаинформация о программе (а точнее, о ее необходимом представлении), такая, как деление на подпрограммы (с аргументами и возвращаемыми значениями), в ряде случаев их имена, детали алгоритмов, схема взаимодействия между программными модулями и т.п. В частности, реинжиниринг путем восстановления предыдущих представлений программы позволяет искать заложенные в них уязвимости [18]. Классическим примером деэволюции представлений является декомпиляция, которая заключается в преобразовании (как правило, с помощью специализированных программных средств) программы из машинного кода в псевдоисходный. Отметим, что под программным кодом понимается любое «задание» программы в строгом (или формализованном) виде, т. е. как классический исходный код, так и бинарный машинный код, а также детализированные блок-схемы, полное и корректное описание архитектуры и т. п.

Далее будет дано представление предлагаемой концепции генетической деэволюции представлений программы с помощью двух следующих схем:

1) (де)эволюции всего множества представлений, описывающей процесс эволюционных изменений программы;
2) генетической деэволюции двух смежных представлений, описывающей получение предыдущего из них по имеющемуся с применением генетических алгоритмов.

Схемы (де)эволюции представлений

Исходя из введенной онтологической модели и ее наложения на практику, логику (де)эволюции представлений программы можно представить в виде следующей схемы (см. рис. 2); используются следующие обозначения: $Form_X$ и $Content_X$ – форма и содержание X-го представления (т.е. имеющего идентификатор X).

Дадим ряд достаточно очевидных пояснений к схеме (де)эволюции представлений (см. рис. 2). Переход от текущего представления (X-1 и X) к следующему представлению (X и X+1) осуществляется в процессе эволюции (согласно схеме жизненного цикла) с помощью некоторого способа синтеза (X-1→X и X→X+1); последним могут быть как программные решения (классические компиляторы и ассемблеры), так и экспертные действия (создание архитектуры, перевод ее в алгоритмы,

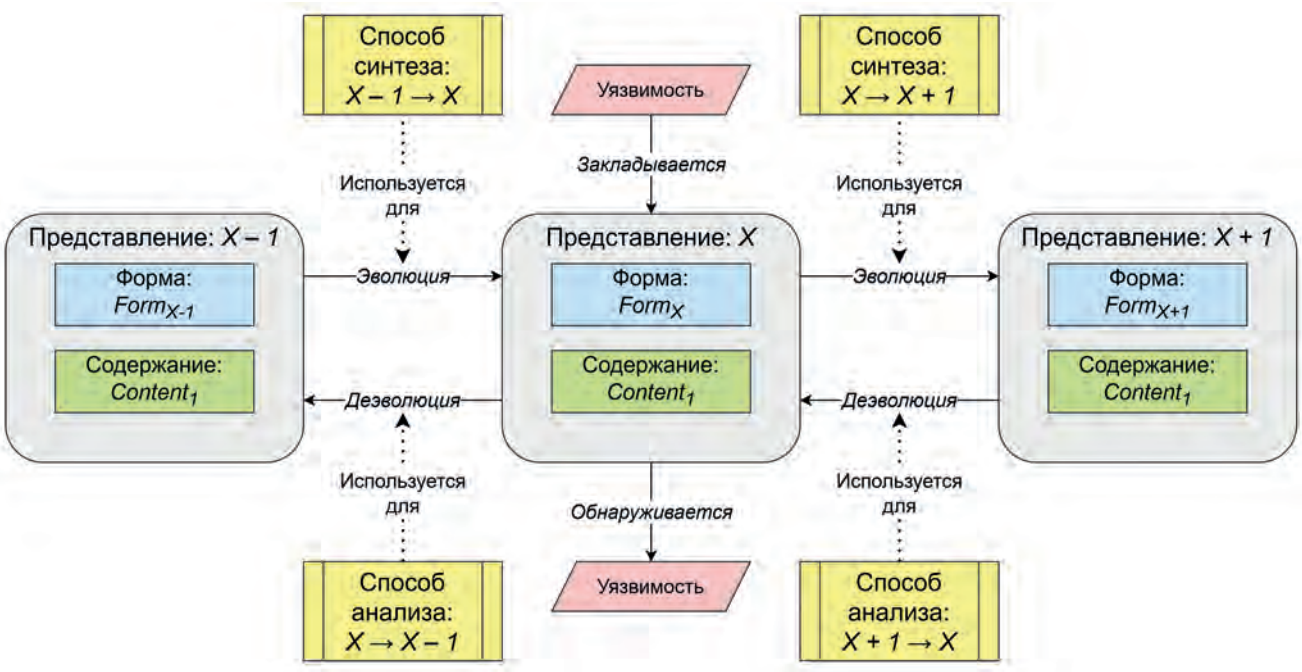


Рис. 2. Общая схема (де)эволюции представлений программы

кодирование программы), а также их замена на генеративный искусственный интеллект [19, 20]. При этом содержание представления не должно меняться, поскольку оно определяет функционал программы, который в конечном представлении должен соответствовать задуманному изначально – таким образом все содержания тождественны созданному в 1-м представлении: $\forall X{:}\,Content_X{=}Content_1$. Форма представления же наоборот подвергается изменениям – от человеко-ориентированного (понятного эксперту и не подходящего для выполнения ЦПУ) к машинно-ориентированному (применимого только для непосредственного выполнения автоматом) – таким образом, формы представлений неидентичны ($\forall X,Y,X{\neq}Y{:}\,Form_X{\neq}Form_Y$), поскольку претерпевают последовательные изменения ($Form_{X\text{-}1}{\rightarrow}Form_X{\rightarrow}Form_{X+1}$). Изменение содержания (случайное или намеренное) в некотором представлении означает появление уязвимостей в программе (т.е. $\exists X,Y{:}\,Content_X{\neq}Content_Y$); поиск же уязвимостей наиболее эффективен именно в том представлении, в котором она была внесена, поскольку в результате эволюции форма уязвимости с объективной неизбежностью также будет видоизменяться. Обратный же процесс получения предыдущих представлений из исходного ($X{+}1{\rightarrow}X$ *и* $X{\rightarrow}X{-}1$) является их деэволюцией. Как было указано выше, данный процесс является наиболее сложным, и именно на качественно новый подход к его реализации и направлено настоящее исследование.

Все процессы, приведенные на схеме, могут быть записаны следующим формальным образом.

Каждое представление программы (*R*, аббр. от англ. ***Representation***) является совокупностью формы (перев. на англ. ***Form***) и содержания (перев. на англ. ***Content***):

$$R \equiv \langle Form, Content \rangle,$$

что также может быть записано через отдельные компоненты R (в случае конкретного представления указывается его идентификатор в нижнем левом индексе):

$$\begin{cases} R^{Form} \equiv Form \\ R^{Content} \equiv Content \end{cases}.$$

Процесс эволюции представлений программы может быть записан как:

$$R_x \rightrightarrows R_{x+1},$$

где «$\rightrightarrows$» – операция эволюции; x – идентификатор текущего представления; а $x{+}1$ – последующего (соответственно, $x{-}1$ – идентификатор предыдущего представления).

Автором выделено несколько типов представлений (перев. на англ. ***Type***), которые используются при классическом инжиниринге программ [5], выполняемых напрямую на ЦПУ (т.е. без применения виртуальных машин [21]):

$$Type^R \in \left\{ \begin{matrix} Type^R_{Идея}, Type^R_{КонцептуальнаяМодель} \\ Type^R_{Архитектура}, Type^R_{Алгоритмы} \\ Type^R_{ИсходныйКод}, Type^R_{МашинныйКод} \end{matrix} \right\},$$

где $Type^R$ – тип представления R, $Type^R_{\dots}$ – идентификатор типа, соответствующий указанному в нижнем индексе представлению; представление ассемблерного кода (и его тип) будет опущено, поскольку оно достаточно хорошо преобразуется в/из машинного и может не учитываться.

Аналогичным образом, процесс деэволюции представлений можно записать следующим образом:

$$R_{x\text{-}1} \leftleftarrows R_x.$$

где «$\leftleftarrows$» – операция деэволюции.

В процессе (де)эволюции происходит лишь изменение формы представления программы, а ее содержание (естественно, в случае неизменности функционала программы – отсутствия привнесенных уязвимостей) остается неизменным, т.е.

$$\forall x,y{:}\,x{\neq}y{:}\begin{cases} R^{Form}_x \neq R^{Form}_y \\ R^{Content}_x = R^{Content}_y \end{cases},$$

где x и y – два различных идентификатора представлений; соответственно, если идентификаторы будут одинаковыми, то и представления программы полностью совпадут.

Для осуществления эволюции ($\rightrightarrows$) представлений применяется соответствующий способ (W, аббр. от англ. ***Way***, перевод на русс. Путь); таким образом, саму операцию можно представить (с помощью символа «:») в виде функции:

$$\rightrightarrows{:}R_{x+1} = E\,(R_x, W_x),$$

где W_x – способ синтеза нового представления из x-го.

Обратную же к эволюции операцию, а именно – деэволюцию ($\leftleftarrows$), можно записать аналогичным образом с использованием диакритического знака отрицания (черты над символом):

$$\leftleftarrows{:}R_{x\text{-}1} = \bar{E}\,(R_x, \bar{W}_x),$$

где $\bar{W}$ – обратный к синтезу способ получения предыдущего представления из x-го путем анализа.

Внесение уязвимостей (V, аббр. от англ. ***Vulnerability***), меняющее, как указывалось, содержание представления программы (без перехода к новой форме), можно записать как:

$$\begin{cases} R_x \rightarrow \widehat{R_x} \\ R^{Form}_x = \widehat{R^{Form}_x} \\ V = \widehat{R^{Content}_x} \setminus R^{Content'}_x \\ V \notin \emptyset \end{cases},$$

где «$\rightarrow$» – операция преобразования представления; R^{Form}_x и $R^{Content}_x$ – форма и содержание -го представления; «$\widehat{\square}$» – диакритический знак для обозначения

сущности (в данном случае, представление) с уязвимостью; «\» – оператор разности двух множеств (в данном случае, содержания представления с и без уязвимости); $\emptyset$ – пустое множество (в данном случае, уязвимостей).

Таким образом, описание внесения уязвимости представляет собой систему следующих уравнений:

1) преобразование безопасного представления в имеющее уязвимость;
2) равенство форм этих представлений;
3) определение уязвимости, как различия содержаний этих представлений;
4) указание на существование уязвимости (т.е. ее отличие от пустого множества).

Аналогичным образом, обнаружение уязвимости можно записать в виде функции ее детектирования (D, аббр. от англ. $Detection$):

$$V = D_x(\widehat{R_x}),$$

где D_x – операция детектирования уязвимости в небезопасном представлении $\widehat{R_x}$, работающая с x-ым представлением. Естественно, реинжиниринг и поиск уязвимостей существенно усложнится в случае применения защит от анализа кода [22].

Продолжение следует...

Литература

1. *Израилов К. Е. Методология реверс-инжиниринга машинного кода. Часть 1. Подготовка объекта исследования. Труды учебных заведений связи. 2023. Т. 9. № 5. С. 79–90. DOI: 10.31854/1813-324X-2023-9-5-79-90*
2. *Bhardwaj V., Kukreja V., Sharma C., Kansal I., Popali R. Reverse Engineering-A Method for Analyzing Malicious Code Behavior // In proceedings of the International Conference on Advances in Computing, Communication, and Control (Mumbai, India, 2021, 03-04 December 2021). PP. 1–5. DOI: 10.1109/ICAC353642.2021.9697150*
3. *Mauthe N., Kargén U., Shahmehri N. A Large-Scale Empirical Study of Android App Decompilation // In proceedings of the IEEE International Conference on Software Analysis, Evolution and Reengineering (Honolulu, HI, USA, 09-12 March 2021). 2021. PP. 400–410. DOI: 10.1109/SANER50967.2021.00044*
4. *Borrello P., Easdon C., Schwarzl M., Czerny R., Schwarz M. CustomProcessingUnit: Reverse Engineering and Customization of Intel Microcode // In proceedings of the IEEE Security and Privacy Workshops (San Francisco, CA, USA, 25-25 May 2023). 2023. PP. 285–297. DOI: 10.1109/SPW59333.2023.00031*
5. *Израилов К. Е. Моделирование программы с уязвимостями с позиции эволюции ее представлений. Часть 1. Схема жизненного цикла // Труды учебных заведений связи. 2023. Т. 9. № 1. С. 75–93. DOI:10.31854/1813-324X-2023-9-1-75-93*
6. *Израилов К. Е. Моделирование программы с уязвимостями с позиции эволюции ее представлений. Часть 2. Аналитическая модель и эксперимент // Труды учебных заведений связи. 2023. Т. 9. № 2. С. 95–111. DOI:10.31854/1813-324X-2023-9-2-95-111*
7. *Самарин Н. Н. Модель безопасного функционирования программного обеспечения, формализующая контроль использования памяти и обращений к ней процессора // Наукоемкие технологии в космических исследованиях Земли. 2021. Т. 13. № 1. С. 68–79.*
8. *Язов Ю. К., Соловьев С. В. Методология оценки эффективности защиты информации в информационных системах от несанкционированного доступа / Санкт-Петербург: Издательство «Наукоемкие технологии». 2023. 258 с.*
9. *Афанасов А. К., Цой А. И. Извлечение данных Android-приложения WhatsApp // Процессы управления и устойчивость. 2021. Т. 8. № 1. С. 246–253.*
10. *Полонский А. М. Импортозамещение программного обеспечения и организация обучения студентов с использованием отечественного или свободного программного обеспечения // Актуальные проблемы экономики и управления. 2022. № 2 (34). С. 65–82.*
11. *Исаев Р. А. Проблемы и перспективы отечественного аналитического программного обеспечения в условиях реализации программ импортозамещения // Промышленные АСУ и контроллеры. 2020. № 11. С. 10–22.*
12. *Шарков И. В. Метод восстановления протокольных автоматов по бинарному коду // Труды Института системного программирования РАН. 2022. Т. 34. № 5. С. 43–62.*
13. *Cao K., Leach K. Revisiting Deep Learning for Variable Type Recovery // In proceedings of the IEEE/ACM 31st International Conference on Program Comprehension (Melbourne, Australia, 15-16 May 2023). 2023. PP. 275–279. DOI: 10.1109/ICPC58990.2023.00042*
14. *Кусаинов А. Р., Глазырина Н. С. Обзор инструментов статического анализа программного кода // Colloquium-Journal. 2020. № 32–1(84). С. 48–52.*
15. *Xu Z., Wen C., Qin S. Type Learning for Binaries and Its Applications // In proceedings of the IEEE Transactions on Reliability. 2019. Vol. 68, No. 3. PP. 893–912. DOI: 10.1109/TR.2018.2884143*
16. *Rani P., Birrer M., Panichella S., Ghafari M., Nierstrasz O. What Do Developers Discuss about Code Comments? // In proceedings of the IEEE 21st International Working Conference on Source Code Analysis and Manipulation (Luxembourg, 27-28 September 2021). 2021. PP. 153–164. DOI: 10.1109/SCAM52516.2021.00027*
17. *Израилов К. Е., Татарникова И. М. Подход к анализу безопасности программного кода с позиции его формы и содержания // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО-2019): сборник научных статей VIII Международной научно-технической и научно-методической конференции (Санкт-Петербург, 27-28 февраля 2019 г.). 2019. С. 462–467.*
18. *Фомин А. И., Хапилина Д. А., Горлищев И. А., Олимпиенко К. В. Инженерный анализ наличия программных закладок в программном обеспечении при отсутствии исходных кодов // Научная мысль. 2019. Т. 7. № 1 (31). С. 123–126.*
19. *Саколик А. ChatGPT и разработка программного обеспечения // БИТ. Бизнес & Информационные технологии. 2023. № 2 (125). С. 38–41.*
20. *Ebert C., Louridas P. Generative AI for Software Practitioners // IEEE Software. 2023. Vol. 40. No. 4. PP. 30–38. DOI: 10.1109/MS.2023.3265877*
21. *Majidha Fathima K. M., Santhiyakumari N. A Survey on Evolution of Cloud Technology and Virtualization // In proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (Tirunelveli, India, 04-06 February 2021). 2021. PP. 428–433. DOI: 10.1109/ICICV50876.2021.9388639.*
22. *Маркин Д. О., Макеев С. М. Система защиты терминальных программ от анализа на основе виртуализации исполняемого кода // Вопросы кибербезопасности. 2020. № 1 (35). С. 29–41. DOI: 10.21681/2311-3456-2020-01-29-4*

МЕТОД АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ЦИФРОВЫХ ОТПЕЧАТКОВ TLS-ПРОТОКОЛА

Ишкуватов С. М.[1], Бегаев А. Н.[2], Комаров И. И.[3]

DOI: 10.21681/2311-3456-2024-1-67-74

Цель исследования: *разработка метода классификации цифровых отпечатков TLS-протокола, обеспечивающего их автоматическое соотнесение с одной из известных групп или принятие решения об обнаружении новой реализации протокола.*

Методы исследования *базируются на положениях теории топологии, теории автоматов, теории множеств; использовании методов автоматической кластеризации, проведения натурного эксперимента и обработки экспериментальных данных.*

Результат: *мониторинг трафика на границе контролируемой зоны телекоммуникационной сети является основной составляющей обеспечения кибербезопасности. Одним из традиционных подходов для решения этой задачи является использование цифровых отпечатков как устройств, так и программной реализации телекоммуникационных протоколов.*

Несмотря на богатую историю развития методов автоматического определения реализации протокола на основе анализа цифровых отпечатков, эта задача в полной мере ещё не решена ввиду изменчивости как самих протоколов, так и телекоммуникационной инфраструктуры, определяющих вариативность конечных формы и значения соответствующего цифрового отпечатка.

В работе предлагается метод автоматической классификации цифровых отпечатков TLS-протокола, базирующийся на формальной оценке близости вариативных форм представления цифровых отпечатков и устойчивый к модификации их значений; приводятся данные по степени влияния значения порога близости на ошибки первого и второго рода в процессе кластеризации.

Полученные результаты в первую очередь ориентированы на применение в системах мониторинга трафика, но могут быть использованы и для решения других задач кибербезопасности.

Научная новизна *результатов определяется совокупностью авторских решений, связанных с обоснованием, введением и применением метрики для оценки близости цифровых отпечатков телекоммуникационных протоколов, устойчивой к модификации цифровых отпечатков клиентской реализации TLS-протокола, а также доказательным подтверждением реализуемости и получением значений показателей качества функционирования метода автоматической классификации цифровых отпечатков реализаций протокола TLS, применённого к известным базам данных цифровых отпечатков.*

Вклад авторов: *Ишкуватов С. М. – разработка метода автоматической классификации цифровых отпечатков протоколов, подготовка исходных данных, проведение эксперимента и визуализация результатов; Бегаев А. Н. – анализ опыта и перспективных сценариев применения периметровых систем мониторинга трафика, определение требований и ограничений исследования; Комаров И. И. – определение научно-методического аппарата и подходов к оценке близости цифровых отпечатков, разработка плана исследования.*

Ключевые слова: *кибербезопасность, мониторинг трафика, коммуникационный протокол, инцидент информационной безопасности, модель пассивного наблюдателя, вектор информативных признаков, мера близости, кластеризация, показатели качества.*

1 Ишкуватов Сергей Маратович, аспирант факультета безопасности информационных технологий, Университет ИТМО, Санкт-Петербург, Россия. E-mail: sysroot0@gmail.com

2 Бегаев Алексей Николаевич, кандидат технических наук, генеральный директор АО «Эшелон – Северо-Запад», Санкт-Петербург, Россия. E-mail: begaev@mail.ru

3 Комаров Игорь Иванович, кандидат физико-математических наук, доцент, доцент факультета безопасности информационных технологий, Университет ИТМО, Санкт-Петербург, Россия. E-mail: i_krov@mail.ru

THE AUTOMATIC METHOD OF TLS PROTOCOL DIGITAL FINGERPRINTS CLASSIFICATION

Ishkuvatov S. M.[4], Begaev A. N.[5], Komarov I. I.[6]

The purpose of the study *is to develop a method for classifying digital fingerprints of the TLS protocol, ensuring their automatic correlation with one of the known groups or making a decision on the discovery of a new protocol implementation.*

The research methods *are based on the principles of topology theory, automata theory, set theory, the use of automatic clustering methods, full-scale experiments, and experimental data processing.*

Results: *traffic monitoring of the telecommunications network-controlled zone border is a key component of ensuring cyber security. One of the traditional approaches to solving this problem is using of digital fingerprints (DF) of both devices and software implementation of telecommunication protocols. Despite the rich history of development automatically determining the protocol implementation methods based on the analysis of DF, this task has not yet been fully solved due to the variability of both the protocols themselves and the telecommunications infrastructure, which determine the variability of the corresponding DF final shape and value.*

The paper proposes an automatic the TLS protocol's DF classification method, based on a formal proximity assessment of the variable forms of DF and resistant to their values modification; data on the influence degree of the proximity threshold value on first and second kind errors in the clustering process are presented.

The results obtained are primarily focused on application in traffic monitoring systems but can also be used to solve other cybersecurity tasks.

Scientific novelty *is determined by a set of author's solutions related to the justification, introduction and application the telecommunication protocols DF proximity assessment metrics that are resistant to the TLS protocol's client implementations modifications, as well as evidence-based confirmation of the feasibility and obtaining the quality indicators values of the automatic classification DF TLS protocol's implementations method functioning applied to known DF databases.*

Keywords: *cybersecurity, informative features vector, digital fingerprint, proximity measure, clustering, digital fingerprint database.*

Введение

Мониторинг и глубокий анализ сетевого трафика на границах сетей является важной составляющей обеспечения кибербезопасности, позволяющий выявлять факты использования небезопасных протоколов, сетевые атаки и обнаруживать иные проблемы информационной безопасности (ИБ).

Одним из традиционных механизмов, используемых системами обнаружения вторжений, является фильтрование трафика по шаблонам – заранее подготовленному списку правил. Однако этот подход опирается на ретроспективный анализ и не позволяет выявлять не описанные ранее угрозы.

Сложность задачи *аналитического* анализа сетевого трафика определяется его имманентной изменчивостью и трудностью формальной интерпретации корректности, связанной как с естественным изменением конфигурации программно-аппаратных средств в процессе развития информационной системы, так и с целенаправленным противодействиям сетевым угрозам [1, 2].

Более того влиятельные транснациональные игроки телекоммуникационной отрасли[7] предпринимают специальные усилия по «запутыванию» протоколов для противодействия национальным цензурам или нежелательным для них способам использования ресурсов.

Промежуточным направлением между «шаблонным» и аналитическим анализом протоколов является подход, основанный на анализе цифровых отпечатков (ЦО), под которыми понимается набор параметров, характеризующий тот или иной протокол, а также позволяющий строить гипотезы относительно реализаций конкретного протокола. Некоторые подходы, связанные с использованием ЦО не только

4 Sergei M. Ishkuvatov, Ph.D. student, Faculty of Information Technology Security, ITMO University, St. Petersburg, Russia. E-mail: hieule250715@gmail.com
5 Alexey N. Begaev, Ph.D., CEO of JSC North-West Echelon, St. Petersburg, Russia. E-mail: begaev@mail.ru
6 Igor I. Komarov, Ph.D., (in Maht.), Associate Professor, Faculty of Information Technology Security, ITMO University, St. Petersburg, Russia. E-mail: i_krov@mail.ru
7 Больше протоколов для шифрования DNSзапросов. — URL: https:// vasexperts.ru/blog/tehnologii/bolsheprotokolovdlyashifrovaniyadnszaprosov/ (дата обращения: 10.10.2023).

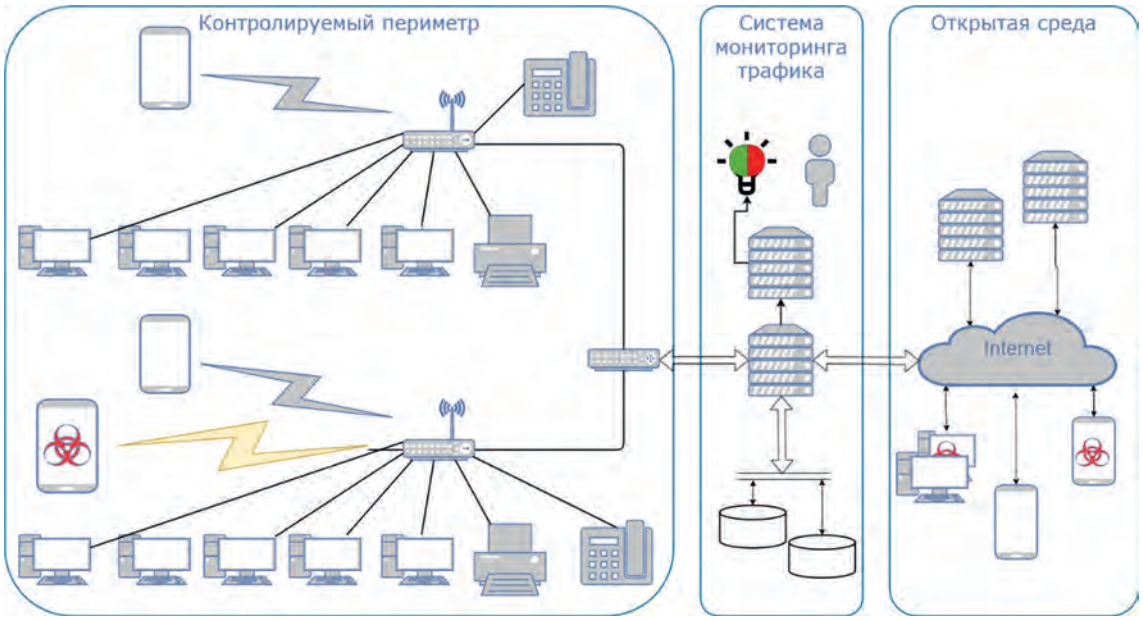


Рис. 1. Расположение систем поведенческого анализа сетевого трафика

программных реализаций, но устройств в целом, доведены[8] до корпоративных стандартов.

В работе предлагается метод, обеспечивающий автоматическую классификацию ЦО программной реализации TLS-протокола с использованием модели пассивного наблюдателя.

Постановка задачи и ограничения

Исследование предполагает (рис. 1) наличие корпоративной защищаемой инфраструктуры, которая находится внутри контролируемого периметра, но вынуждена взаимодействовать с внешней неконтролируемой средой. Взаимодействие осуществляется через систему мониторинга трафика, к которой применена модель «пассивного наблюдателя». Задача системы мониторинга – информирование администратора сети об обнаруженных инцидентах таких как появление новых абонентов/видов сетевых активностей или сокрытия фактов использования запрещённых организацией протоколов. Решение этой задачи осложняется:

- Распространением технологий сокрытия DNS запросов, таких как DNS over HTTPS, DNS over TLS, DNS over QUIC, DNS over CoAP, SecureDNS.
- Проблемами типа DNS leaks[9] – когда запрос локального узла к локальному ресурсу ошибочно перенаправляется глобальному DNS, раскрывая внутреннюю архитектуру сети на всем пути следования пакета [3].
- Расширением применения шифрованного TLS-рукопожатия Encrypted Client Hello (ECH), делающим невозможным[10] определения конечной точку TLS-сессии по значению поля ServerName из запроса клиента и проверки цепочки сертификатов, предоставляемых сервером.

В результате система мониторинга трафика сталкивается с рядом сложностей, в том числе, нарушающих функционирование прикладной системы:

- TLS-сессии станут практически неотличимы друг от друга традиционными методами;
- станет невозможным Sinkholing[11] – получение информации о заражении перенаправлением вредоносного трафика на сервер исследователя;
- ограниченное использование баз данных (БД) ЦО, например на основании дефакто-стандарта JA3[12];
- системы контент фильтрации смогут блокировать ресурсы только в случае явного обращения к запрещённому ресурсу по IPv4 адресу;
- невозможность выборочной блокировки ресурса без блокировки всех ресурсов, использующих эту сеть доставки контента CDN[13].

Преодоление указанных сложностей системы анализа трафика может реализовываться в следующих направлениях:

- поиск новых *информативных признаков* и *подходов к описанию* информативных признаков

8 СТО БР БФБО-1.7-2023 Стандарт Банка России «Безопасность финансовых (банковских) операций. Обеспечение безопасности финансовых сервисов с использованием технологии цифровых отпечатков устройств (принят и введён в действие приказом Банка России от 01.03.2023 N ОД-335)

9 Imana Basileal, Korolova Aleksandra, Heidemann John. Enumerating privacy leaks in DNS data collected above the recursive // NDSS: DNS Privacy Workshop. — 2018.

10 Encrypted Client Hello (ECH): часто задаваемые вопросы. — URL: https://support.mozilla.org/ru/kb/faq-encrypted-client-hello (дата обращения: 20.05. 2023).

11 Sinkholing — URL: https://encyclopedia.kaspersky.ru/glossary/sinkholing/ (дата обращения: 10.10.2023).

12 JA3 — A method for profiling SSL/TLS Clients [Электронный ресурс]. URL: https://github.com/salesforce/ja3 (дата обращения: 19.07.2020).13

13 Peng Gang. CDN: Content distribution network // arXiv preprint cs/0411069. — 2004.

реализаций протоколов для учёта их вариативности и постоянной мимикрии угроз;
- совместный анализ информативных признаков *разных протоколов* и на разных уровнях модели OSI [4];
- поиск *статистических закономерностей* вредоносного трафика, которые могут быть выявлены пассивным наблюдателем даже в случае использования сторонами шифрования.

Предпосылки исследования

Исторически сложилось два основных направления получения данных об устройствах и протоколах сети: активный и пассивный [5]. Несмотря на более широкие возможности активных методов, их применение не всегда возможно.

Исследования, посвящённые выделению и описанию информативных признаков, процедуре *пассивного* получения ЦО[14], а также его использования, в том числе в контексте HTTPS трафика[15], получали развитие по мере развития телекоммуникационных систем.

Результаты, полученные в работах [6 – 7] определили возможность применения ЦО для выявления угроз ИБ и легли в основу механизмов идентификации реализаций TLS-протокола, развиваемых проектами JA3 и JA3S, а также Cisco Mercury. В настоящее время в открытой БД ЦО проекта JA3 используется две формы: исходная полная запись признаков и *md5*-хеш этой полной формы.

Отдельную группу составляют работы, посвящённые проблеме классификации трафика, передаваемого в шифрованной сети [8–13]. Решения демонстрируют хорошие результаты по определению типа трафика на основании анализа нормализованных по времени и размерам распределений длин пакетов, как для отдельных TLS-сессий, так и всего канала VPN.

В работе[16] предлагаются обзор перспективных подходов, в том числе не ограниченных признаками TLS-рукопожатия, таких как цепи Маркова, описывающие сетевое взаимодействие сторон.

Анализ рынка систем анализа трафика позволяет выделить несколько ключевых проектов, характеризующих достигнутый практический уровень.

Характерным представителями систем, базирующихся на проекте JA3/JA3S является продукты Wireshark и Suricata[17] , использующие, в том числе, модели [14, 15].

Известна отечественная система анализа трафика для выявления атак PT Network Attack Discovery[18]. В контексте исследования особый интерес представляет библиотека OsDetectLib[19], которая по описанию разработчиков занимается определением операционных систем TCP-сессий. В репозитории на Github разработчик публикует открытую часть правил детектирования различных видов атак в формате Suricata[20]. Формат правил Suricata также имеет функционал автоматического получения ЦО TLS реализаций в формате JA3/JA3S и возможность описания ЦО TCP/IP. Однако правила, содержащие ЦО различных уровней, являются строгими и не предполагают оценок возможной близости, кроме того, в части случаев ЦО задаётся MD5-хешем, что хоть и делает правила более компактными, препятствует любой проверки на соответствие, кроме строгой.

Способ решения задачи и анализ полученных результатов

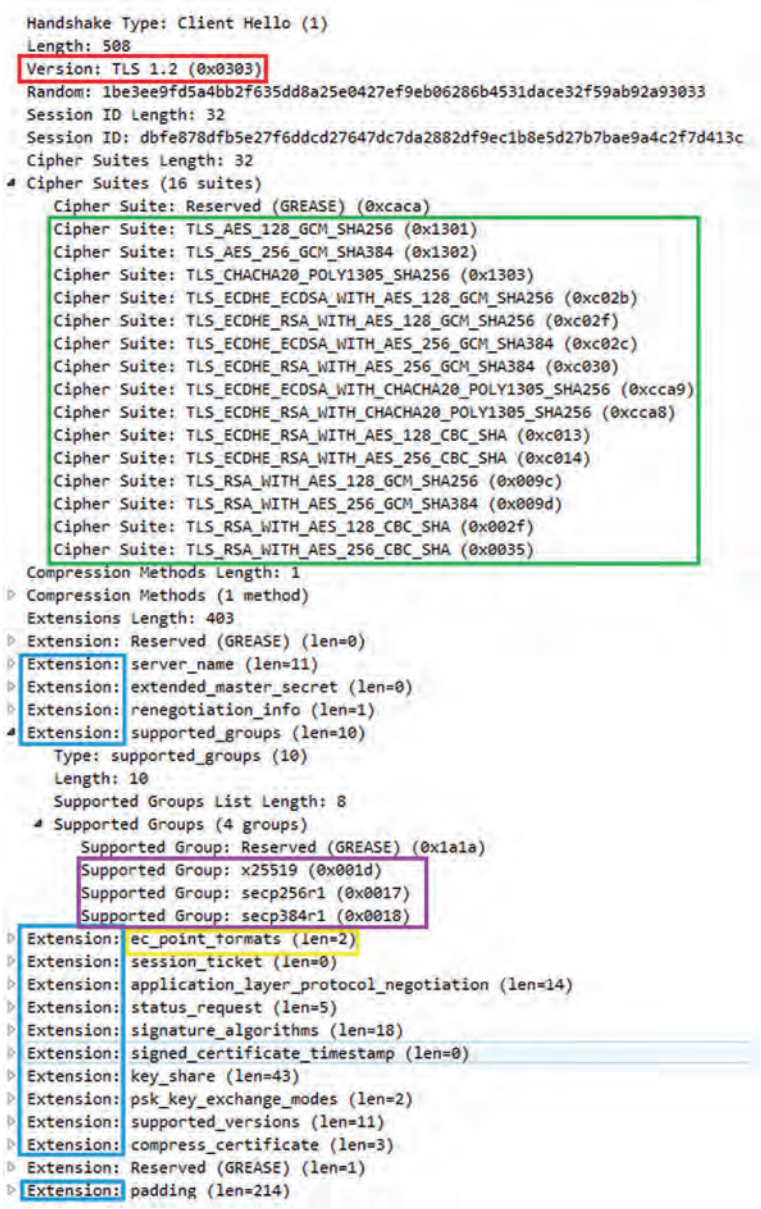
Для решения задачи автоматической классификации ЦО TLS-протоколов должны быть решены следующие частные задачи:
- определена номенклатура информативных признаков, доступных пассивному наблюдателю;
- выбрана единая (псевдоканоническая) форма записи наблюдаемых признаков;
- введена метрика близости ЦО, обеспечивающая формальную оценку расстояния ЦО в многомерном пространстве признаков;
- выполнена подготовка БД ЦО для использования в автоматическом режиме;
- предложен алгоритм кластеризации ЦО.

Под термином ЦО реализации TLS-протокола понимаются параметры, характеризующие именно эту *конкретную реализацию* протокола, именно *конкретную версию библиотеки*, реализующий этот протокол или *группу возможных версий*.

Для решения задач формирования ЦО клиентской реализации TLS-протокола наибольшее распространение в настоящее время получил алгоритм JA3 [16], интересны модели использования альтернативных баз представления признаков – проекта Mercury Cisco[21,22], [17, 18, 19] и LeeBrotherston[23] [20].

Критерием выбора информативных признаков TLS-протокола для использования в работе (рис. 2) являются: 1) возможность определения пассивным

14 Shu G., Lee D. Network protocol system fingerprinting a formal approach // Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications. — IEEE, 2006. — Pp. 1–12.
15 HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting / Martin Husák, Milan Čermák, Tomáš Jirsík, Pavel Čeleda // EURASIP Journal on Information Security. 2016. Vol. 2016. Pp. 1–14.
16 Gancheva Z., Sattler P., Wüstrich L. TLS Fingerprinting Techniques // Network. — 2020. — URL: https://www.net.in.tum.de/fileadmin/TUM/NET/NET-2020-04-1/ NET-2020-04-1_04.pdf (online; accessed: 20.05.2023).
17 Suricata. Observe. Protect. Adapt. — URL: https://suricata.io/ (online; accessed: 10.10.2023).
18 PT Network Attack Discovery. — URL: https://ptsecurity.com/ru-ru/products/network-attack-discovery/ (дата обращения: 20.05.2023).
19 Результаты анализа трафика в 41 компании и новые возможности PT NAD. — URL: https://www.ptsecurity.com/upload/corporate/ru-ru/webinars/ics/PT_NAD_ 18_03.pdf (дата обращения: 20.05.2023).
20 Suricata PT Open Ruleset. — URL: https://github.com/ptresearch/AttackDetection (дата обращения: 20.05.2023).
21 Mercury: network fingerprinting and packet metadata capture - URL:https://github.com/cisco/mercury. (online; accessed: 21.10.2023).
22 Lee brotherston's work – URL: https://github.com/synackpse/tls-fingerprinting (online; accessed: 21.10.2023).
23 Brotherston Lee. Lee brotherston's work. — URL: https://github.com/synackpse/tls-fingerprinting (online; accessed: 20.05.2023).



а) Информативные признаки пакета TLS Client Hello, используемые для формирования ЦО

```
[JA3 Fullstring: 771,4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53,0-23-65281-10-11-35-16-5-13-18-51-45-43-27-21,29-23-24,0]
[JA3: b32309a26951912be7dba376398abc3b]
```

б) Полный и хешированный ЦО в формате JA3, полученные из данных рис. 2.а)

Рис.2. – Формирования ЦО пакета TLS Client Hello

наблюдателем и 2) возможность их получения из распространённых БД ЦО.

Перечень информативных признаков включает:

- номер используемой версии протокола TLS – целое число (выделено красным);
- список поддерживаемых клиентской реализации алгоритмов шифрования Cipher Suites – последовательность 2-байтных символов (выделено зелёным);
- список опциональных параметров TLS Extensions последовательность 2-байтных символов (выделено голубым);
- хронология следования параметров TLS Extensions (формируется динамически);
- EC point formats (выделено жёлтым) – в случае, если этот тип поля присутствует только в одном ЦО, принимать расстояние равным 2;
- список Elliptic Curves – в случае, если этот тип поля присутствует только в одном ЦО, принимать расстояние равным 2.

Несмотря на возможность произвольной авторской модификации возможных форматов хранения ЦО, открывающих новые перспективы автоматиче-

ской обработки, в качестве *псевдоканонической формы записи* ЦО для выбран *полный формат JA3*. Выбор определяется его широким использованием в профессиональном сообществе, активным пополнением базы и поддержкой значительным числом программных продуктов, что упрощает вывод в практическое применение полученных результатов.

Введение метрики для оценки близости двух ЦО протокола предполагает определение метрического пространства [21] и способов обработки каждого из компонентов вектора признаков. Независимо от протокола все возможные признаки, описывающие ЦО можно разделить на следующие типы:

- флаги – булевские атрибуты, характеризующие наличие или отсутствие определённого признака у характеризуемой им реализации протокола;
- константное числовое значение;
- диапазон значений – применим для числовых значений, может определяться формулой, списком или границами интервалов значений;
- последовательность – обычно список мнемонических обозначений параметров с имеющей значение хронологией следования элементов друг за другом – для описания порядка следования и состав опциональных параметров.

В работе [21] предложен метод количественной оценки отличий $\Delta i(a,b)$ каждого из i значений компонента векторов признаков $A=<a_1, a_2...a_n>$ и $B=<b_1, b_2...bn>$. Она может определяться:

- если значения всех типов признаков a и b совпадают $\Delta i(a,b) = \emptyset$;
- в противном случае:
- для числовых констант – как абсолютное значение разности a и b: $\Delta i(a,b) = |(a - b)|$;
- для диапазонов – размер диапазона, образованного пересечением диапазонов $\Delta i(a,b) = M|a_i \cap b_i|$, (где M – мощность множества), возможно с модификацией: $\Delta i(a,b)=\frac{M|a_i \cap b_i|}{M|a_i \cup b_i|}$, обеспечивающей учёт относительной мощности сравниваемых множеств.
- для последовательностей – количественная оценка совпадения состава и порядка следования мнемоник, полученная как расстояние Левенштейна. Алгоритм определения расстояния Левенштейна lev(a,b) для последовательностей a и b с длинами $|a|$ и $|b|$ определяется (1), где $tail$ некоторой строки x – это строка всех символов x, кроме первого, а $x[n]$ – n-ный символ строки x, начиная $\emptyset$.

Доказана применимость настоящего подхода [21], основанная на том, что для подавляющего числа протоколов добавление или удаление параметра из списка, как правило, не приводит к нарушению общей хронологии следования параметров.

$$lev(a,b) = \begin{cases} |a| & \text{, если } |b| = 0 \\ |b| & \text{, если } |a| = 0 \\ lev(tail(a),tail(b)) & \text{, если } a|0| = b|0| \\ 1 + min \begin{cases} lev(tail(a),b) \\ lev(a,tail(b)) \\ lev(tail(a),tail(b)) \end{cases} & \text{, в остальных случаях} \end{cases} \quad (1)$$

Общим расстоянием между двумя отпечатками $LEV(A,B)$ следует считать сумму всех минимальных расстояний Левенштейна всех компонентов вектора признаков:

$$LEV(A,B) = \sum k \cdot \min(lev_i(a_i, b_i)), \quad (2)$$

где i – индекс вектора гиперпространства, в котором вычисляется (1) расстояние Левенштейна $lev_i(a_i, b_i)$; k – весовой коэффициент (коэффициент значимости) значения расстояния по вектору i.

Определение пространства информативных признаков и введение формальной метрики оценки близости ЦО предоставляет возможность автоматической классификации (рис. 3) всех известных наборов ЦО.

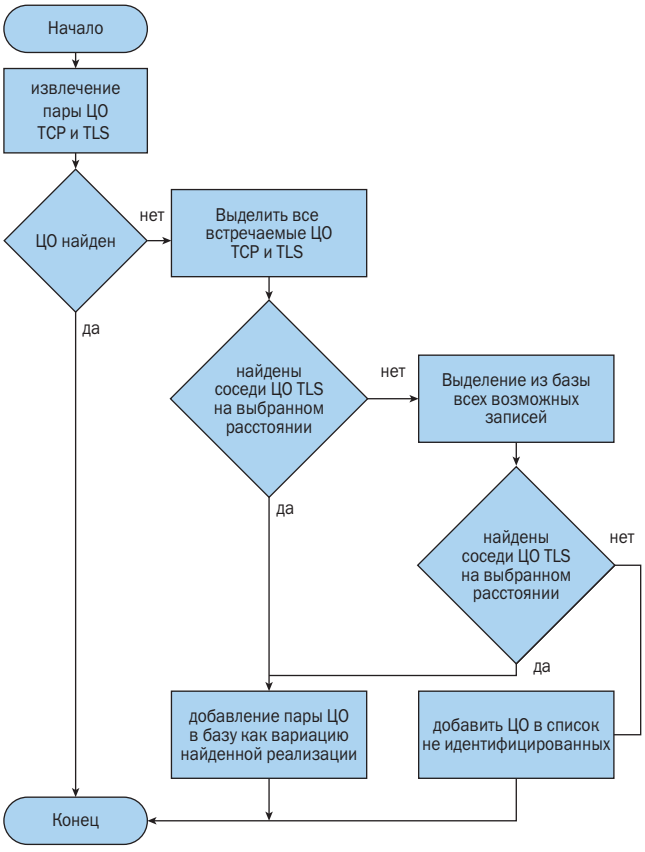


Рис.3. – Алгоритм автоматической классификации ЦО TLS-протокола

В ходе автоматической классификации ЦО реализаций протокола обнаружено явление *смещения центра кластеров родственных реализаций* по мере появления новых версий, как за счёт новых реализаций, так и за счёт использования новых библиотек. Этот факт должен учитываться при реализации подсистем ИБ, базирующихся на предлагаемом методе.

Таблица 1

Показатели качества функционирования метода автоматической классификации

Представление ЦО TLS: ***Источник*** (Полная запись ЦО) *Хеш-значение*	max расстояние близости	Корректно найденные соседи	Ошибки I рода (ложные соседи)	Ошибки II рода (элементы, ошибочно не включённые в класс)
Android Webkit (771,49195-49196-49199-49200-158-159-49161-49162-49171-49172-51-57-50-56-49159-49169-156-157-47-53-5-255,0-11-10-13,14-13-25-11-12-24-9-10-22-23-8-6-7--20-21-4-5-18-19-1-2-3-15-16-17,0-1-2) *f898478e132de326106e9eb8e861c1a2*	6	11	0	443
	20	16	1	438
	30	74	6	380
	50	85	325	369
Tor (769,49162-49172-136-135-57-56-49167-49157-132-53-49159-49161-49169-49171-69-68-51-50-49164-49166-49154-49156-150-65-4-5-47-49160-49170-22-19-49165-49155-65279-10-255,0-11-10,1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22-23-24-25,0-1-2) *581a3c7f54555512b8cd16e87dfe165b*	6	0	1	10
	20	1	1	9
	30	3	5	7
Kaspersky (771,4866-4867-4865-49200-49199-49192-49191-49196-49195-49188-49187-52392-52394-103-107-159-255,0-11-10-35-5-16-22-23-49-13-43-45-51-21,23-24-29,0-1-2) *aa63ca1ce311b0ff100de506d4d9b3ab*	6	20	0	19
	20	23	0	16
	30	24	135	15

Эксперимент и анализ полученных результатов

Эксперименты по автоматической классификация ЦО TLS-протокола проведены в два этапа.

Первый этап предполагает использование классических методов кластеризации для графического представления и визуальной интерпретации результатов. На рис. 4 представлен фрагмент дендрограммы,

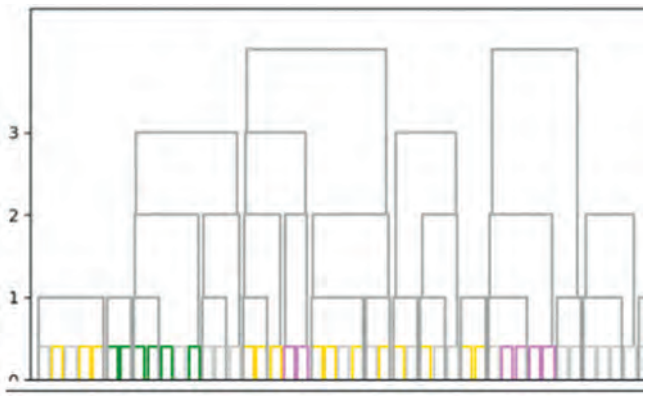


Рис.4. – Фрагмент дендрограммы кластеризации БД ЦО Cisco протокола TLS

где одинаковыми цветами обозначены родственные реализации протоколов, серым цветом обозначены все редко встречающиеся реализации. Такое представление позволяет визуально оценить корректность выбора порогового значения по размерам графов и количествам попаданий элементов разного цвета в один граф. Видно, что на представленном фрагменте наибольшее расстояние $LEV(A,B)$, объединяющее все известные реализации, равно 5.

Второй этап эксперимента нацелен на определение влияния порогового значения близости ЦО. Очевидно, что выбор высокого значения ведёт к увеличению ошибок первого рода (ложное принятие схожести ЦО), что в дальнейшем приведёт к слиянию разных групп ЦО и невозможности их разделения.

Снижение значения порога близости, хоть и ведёт к увеличению ошибок второго рода (ложное предположение о непохожести ЦО), однако вред таких ошибок менее значим, так как он ведёт к дроблению групп ЦО на более мелкие группы, которые при необходимости могут быть объединены на последующих этапах.

По результатам вычислительного эксперимента на текущей[24] БД ЦО Cisco, приведённой в формат JA3 и дополненной информацией из открытых баз[25] [24] получены результаты, представленные в таблице 1.

Полученные зависимости позволяют производить тонкую настройку прикладных систем с учётом степени важности ошибок первого и второго рода, например на основе рисковых моделей, например [22].

Выводы

В работе поставлена и решена задача разработки метода автоматической классификации цифровых отпечатков TLS-протокола. Предлагаемый метод базируется на совокупности ранее полученных результатов, связанных с обоснованием и выбором информативных признаков, введением метрики для оценки близости ЦО протоколов, обработкой открытых БД ЦО и использованием методов кластеризации данных.

Теоретические результаты подтверждены экспериментальным исследованием, в том числе, определяющим степень влияния порога близости ЦО на результат отнесения исследуемого ЦО к известным или новым кластерам.

Предлагаемые результаты ориентированы на применение в системах периметрового мониторинга трафика с использованием модели пассивного наблюдателя, однако могут найти применения и для ряда задач, требующих оценки аутентичности трафика, проходящего через канал.

24 База данных цифровых отпечатков Cisco – URL: https://github.com/cisco/mercury/blob/main/resources/fingerprint_db.json.gz (дата обращения: 20.09.2023).

25 Открытый формат представления цифровых отпечатков ja3 URL: https://github.com/trisulnsm/trisul-scripts/blob/master/lua/frontend_scripts/reassembly/ja3/prints/ja3fingerprint.json (дата обращения: 20.09.2023).

Литература

1. *Ворончихин И. С., Иванов И. И., Максимов Р. В., Соколовский С. П. Маскирование структуры распределённых информационных систем в киберпространстве // Вопросы кибербезопасности. 2019. № 6 (34). – С. 92–101. DOI: 10.21681/2311-3456-2019-6-92-101*
2. *Москвин А. А., Максимов Р. В., Горбачёв А. А. Модель, оптимизация и оценка эффективности применения многоадресных сетевых соединений в условиях сетевой разведки // Вопросы кибербезопасности. 2023. № 3 (55). – С. 13–22.*
3. *Tatang Dennis, Schneider Carl, Holz Thorsten. Largescale analysis of infrastructureleaking DNS servers // Detection of Intrusions and Malware, and Vulnerability Assessment: 16th International Conference, DIMVA 2019, Gothenburg, Sweden, June 19–20, 2019, Proceedings 16 / Springer. — 2019. —Pp. 353–373*
4. *Клименко Т. М., Акжигитов Р. Р. Обзор методов обнаружения распределённых атак типа» отказ в обслуживании» на основе машинного обучения и глубокого обучения //International Journal of Open Information Technologies. – 2023. – Т. 11. – №. 6. – С. 46–66.*
5. *Dangi A., Batra U. TLS Fingerprinting «A Passive Concept of Identification» //Artificial Intelligence and Machine Learning in Healthcare. – Singapore: Springer Nature Singapore, 2023. – С. 95-116.*
6. *Althouse J., Atkinson J., Atkins J. TLS fingerprinting with JA3 and JA3S //Salesforce. – 2019.*
7. *Rana S., Garg U., Gupta N. Intelligent Traffic Monitoring System Based on Internet of Things //2021 International Conference on Computational Performance Evaluation (ComPE). – IEEE, 2021. – С. 513–518.*
8. *Полянская М. С. Анализ подходов к обнаружению атак в зашифрованном трафике // Современные информационные технологии и ИТ-образование. 2021. Т. 17, No 4. С. 922–931. DOI: https://doi.org/10.25559/SITITO.17.202104.922-931*
9. *Ali Rasteh, Florian Delpech, Carlos AguilarMelchor et al. Encrypted internet traffic classification using a supervised spiking neural network // Neurocomputing. — 2022. — Vol. 503. — Pp. 272–282.*
10. *Gupta Neha, Jindal Vinita, Bedi Punam. Encrypted traffic classification using extreme gradient boosting algorithm // International Conference on Innova tive Computing and Communications: Proceedings of ICICC 2021, Volume 3 / Springer. — 2022. — Pp. 225–232.*
11. *Islam Faiz Ul, Liu Guangjie, Liu Weiwei. Identifying VoIP traffic in VPN tunnel via flow spatiotemporal features // Mathematical Biosciences and Engineering. — 2020. — Vol. 17, no. 5. — Pp. 4747–4772.*
12. *Islam F. U. et al. VoIP traffic detection in tunneled and anonymous networks using deep learning // IEEE Access. – 2021. – Т. 9. – С. 59783–59799.*
13. *Li K., Cui B. Malicious Encrypted Traffic Identification Based on Four-Tuple Feature and Deep Learning // Innovative Mobile and Internet Services in Ubiquitous Computing: Proceedings of the 15th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2021). – Springer International Publishing, 2022. – С. 199–208.*
14. *Sismis L., Korenek J. Analysis of TLS Prefiltering for IDS Acceleration //International Conference on Passive and Active Network Measurement. – Cham: Springer Nature Switzerland, 2023. – С. 85–109.*
15. *Deri L., Fusco F. Using Deep Packet Inspection in CyberTraffic Analysis //2021 IEEE International Conference on Cyber Security and Resilience (CSR). – IEEE, 2021. – С. 89–94.*
16. *Anderson Blake, McGrew David. Accurate TLS fingerprinting using destination context and knowledge bases // arXiv preprint arXiv:2009.01939. — 2020.*
17. *Anderson B., McGrew D. Tls beyond the browser: Combining end host and network data to understand application behavior //Proceedings of the Internet Measurement Conference. – 2019. – С. 379–392.*
18. *Varmarken J. et al. FingerprinTV: Fingerprinting Smart TV Apps //Proceedings on Privacy Enhancing Technologies (PoPETs). – 2022. – Т. 2022. – №. 3. – С. 606–629.*
19. *Kim H. et al. Revisiting TLS-Encrypted Traffic Fingerprinting Methods for Malware Family Classification //2022 13th International Conference on Information and Communication Technology Convergence (ICTC). – IEEE, 2022. – С. 1273–1278.*
20. *Heino J. et al. On usability of hash fingerprinting for endpoint application identification // 2022 IEEE International Conference on Cyber Security and Resilience (CSR). – IEEE, 2022. – С. 38–43.*
21. *Ишкуватов С. М., Швед В. Г., Филькова И. А. Метод оценки близости цифровых отпечатков реализаций протоколов // Информационно-методический журнал «Защита информации. Инсайд». — 2022. — № 2. — С. 29–33.*
22. *Беляев Е. А., Емельянова О. А., Лившиц И. И. Анализ методик оценки рисков информационной безопасности кредитно-финансовых организаций // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 3. С. 437–441. DOI: 10.17586/2226-1494-2021-21-3-437-441*

ФОРМИРОВАНИЕ УЯЗВИМОГО УЗЛА «ADOBE COLDFUSION DESERIALIZATION OF UNTRUSTED DATA VULNERABILITY»

Конев А. А.[1], Репкин В. С.[2], Семёнов Г. Ю.[3], Сермавкин Н.И.[4]

DOI: 10.21681/2311-3456-2024-1-75-81

Цель исследования: *разработка уязвимого узла, что включает в себя анализ исследуемой уязвимости, реализацию её автоматизированной эксплуатации, формализацию процесса атаки, описание способов обнаружения, а также мер защиты.*

Методы исследования: *системный анализ, формализация процесса эксплуатации уязвимости с помощью методологии моделирования Meta Attack Language (MAL).*

Результат исследования: *в данной научной публикации представлен подробный анализ уязвимости «Adobe ColdFusion Deserialization of Untrusted Data Vulnerability» (CVE-2023-26360) формальное описание процесса ее эксплуатации с использованием MAL. Работа включает в себя описание структуры формируемого уязвимого узла и потенциальных угроз. Кроме того, статья представляет практический сценарий автоматизированной атаки, осуществляемой с использованием Python и фреймворка Metasploit, который может быть использован специалистами для определения защищенности собственной информационной системы. На основе проведенного исследования, в работе приводятся меры защиты и рекомендации для снижения риска эксплуатации уязвимости, включая установку обновлений безопасности и отключение компонентов, представляющих уязвимость.*

Практическая значимость: *результаты исследования можно использовать при создании и формализации сценариев атак, отмеченные меры защиты и детальное описание уязвимости могут быть использованы для обеспечения безопасной разработки на языке ColdFusion, представленный в работе код может быть применен в тестировании систем на проникновение. В данной научной статье не только анализируется уязвимость, но и демонстрируются все шаги её эксплуатации, что позволяет разработать более эффективные методы защиты информационных систем от подобных атак.*

Вклад авторов: *Конев А. А. выполнил постановку задачи и определил методы исследования. Сермавкин Н. И. разработал и настроил сетевую инфраструктуру, провел анализ уязвимости. Семенов Г. Ю. реализовал атаку, эксплуатирующую уязвимость с помощью Metasploit Framework, формализовал атаку с помощью MAL. Репкин В. С. реализовал автоматизированную атаку с помощью Pymetasploit, определил меры защиты.*

Ключевые слова: *информационная безопасность, обучение специалистов, автоматизированная эксплуатация, меры защиты, тестирование на проникновение, имитация атаки, киберполигон, Metasploit, Remote Code Execution, Meta Attack Language.*

FORMATION OF VULNERABLE NODE «ADOBE COLDFUSION DESERIALIZATION OF UNTRUSTED DATA VULNERABILITY»

Konev A. A.[5], Repkin V. S.[6], Semenov G. Yu.[7], Sermavkin N. I.[8]

The purpose of the article: *development of a vulnerable node, which includes the analysis of the vulnerability under study, implementation of its automated exploitation, formalization of the attack process, description of detection methods, as well as protection measures.*

1 Конев Антон Александрович, кандидат технических наук, доцент, ФГБОУ ВО «Томский государственный университет систем управления и радиоэлектроники» (ТУСУР), г. Томск, Россия. E-mail: kaa@fb.tusur.ru

2 Репкин Владимир Сергеевич, техник, Центр компетенций Национальной технологической инициативы «Технологии доверенного взаимодействия» (ЦК НТИ ТДВ), г. Томск, Россия. E-mail: repkin_vova@mail.ru

3 Семёнов Григорий Юрьевич, техник, Центр компетенций Национальной технологической инициативы «Технологии доверенного взаимодействия» (ЦК НТИ ТДВ), г. Томск, Россия. E-mail: semenov.g.749-1@e.tusur.ru

4 Сермавкин Никита Игоревич, техник, Центр компетенций Национальной технологической инициативы «Технологии доверенного взаимодействия» (ЦК НТИ ТДВ), г. Томск, Россия. E-mail: iis.vseverske@mail.ru

5 Anton A. Konev, Ph.D., Associate Professor, Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, Russia. E-mail: kaa@fb.tusur.ru

6 Vladimir S. Repkin, Technician, Center of Competences of the National Technological Initiative «Trusted Interaction Technologies», Tomsk, Russia. E-mail: repkin_vova@mail.ru

7 Grigory Y. Semenov, Technician, Center of Competences of the National Technological Initiative «Trusted Interaction Technologies», Tomsk, Russia. E-mail: semenov.g.749-1@e.tusur.ru

8 Nikita I. Sermavkin, Technician, Center of Competences of the National Technological Initiative «Trusted Interaction Technologies», Tomsk, Russia. E-mail: iis.vseverske@mail.ru

Research method: *system analysis, formalization of the vulnerability exploitation process using Meta Attack Language (MAL) modeling methodology.*

The result: *this scientific publication presents a detailed analysis of the vulnerability «Adobe ColdFusion Deserialization of Untrusted Data Vulnerability» (CVE-2023-26360) and a formal description of the process of its exploitation using MAL. The paper includes a description of the structure of the vulnerability node being formed and the potential threats. In addition, the paper presents a practical scenario of an automated attack carried out using Python and the Metasploit framework, which can be used by specialists to determine the security of their own information system. Based on the research conducted, the paper provides protective measures and recommendations to reduce the risk of vulnerability exploitation, including installing security updates and disabling components that present the vulnerability.*

Practical significance: *the results of the study can be used in the creation and formalization of attack scenarios, the noted protection measures and a detailed description of the vulnerability can be used to ensure secure development in the ColdFusion language, the code presented in the paper can be applied in penetration testing of systems. This research paper not only analyzes the vulnerability, but also demonstrates all the steps of its exploitation, which allows us to develop more effective methods of protecting information systems from such attacks.*

Keywords: *information security, specialist training, automated operation, protection measures, penetration testing, attack simulation, cyberpolygon, Metasploit, Remote Code Execution, Meta Attack Language.*

Введение

На данный момент в области кибербезопасности существует серьезная проблема, связанная с нехваткой практических навыков в обучении специалистов[9] по выявлению и предотвращению инцидентов информационной безопасности [1, 2]. Обычно учебные программы ориентированы на теоретическое обучение, что, безусловно, важно в данной сфере, но недостаточно для эффективного противодействия современным кибератакам. Эта проблема актуальна как для начинающих, так и для более опытных специалистов, поскольку постоянное обновление практических и теоретических знаний, следование последним тенденциям и готовность к новым векторам атак имеют ключевое значение.

Для решения этой задачи существует киберполигон Ampire[10], который представляет собой учебно-тренировочную площадку для проведения массовых киберучений. Основой этой платформы являются имитации реальных кибератак. Ampire позволяет специалистам разрабатывать и оттачивать свои навыки в условиях, максимально приближенных к реальным угрозам и атакам, что делает его важным инструментом для обучения и поддержания актуальности знаний в области кибербезопасности. Для проведения тренировок используются уязвимые узлы. Уязвимый узел – это компьютерная система, которая состоит из двух виртуальных машин, связанных в сети. Одна из этих машин является хостом злоумышленника и используется для проведения автоматизированных атак на вторую машину. Под атакой понимается эксплуатация имеющейся уязвимости в предустановленном программном обеспечении, сервисах или операционной системе второй виртуальной машины. Уязвимый узел используется в обучении и тестировании в области кибербезопасности с целью оценки и усовершенствования защиты информационных систем.

Увеличение количества таких узлов способствует повышению разнообразия изучаемых уязвимостей, что в свою очередь обогащает образовательный процесс и делает его более реалистичным. Это помогает специалистам в области кибербезопасности получать более широкий опыт и быть готовым к новейшим вызовам в области кибербезопасности.

Обзор исследований

Современным методам подготовки специалистов по информационной безопасности уделяется много внимания в научных и исследовательских работах [3–5]. В статье [3] предлагается оборудовать профильные высшие учебные заведения специальными пентест-лабораториями. В работах [4,5] подтверждается важность наличия у специалиста по информационной безопасности готовности принимать активные действия по нейтрализации угрозы. Таким образом, можно понять, что ключевой задачей в области обучения профильных специалистов является получение практических навыков.

Невозможность работы начинающих или еще обучающихся специалистов с реальными инцидентами информационной безопасности обусловила появление новых методов обучения, а именно получения

9 Теории недостаточно: о важности практических навыков при обучении сотрудников кибербезопасности [Электронный ресурс]. URL: https://www.anti-malware.ru/analytics/Threats_Analysis/importance-of-practical-skills-in-cybersecurity-training (дата обращения: 16.10.2023).

10 Киберполигон Ampire [Электронный ресурс]. URL: https://amonitoring.ru/ampire-po/ (дата обращения: 29.09.2023).

практических навыков с помощью киберполигонов и соревнований типа Capture the Flag (CTF) [6–8]. В работах [7, 8] целью исследования является изучение соревнований CTF и виртуальных лабораторий как инструментов для получения практических навыков и прикладных умений. В статье [6] выявляется преимущество использования киберполигонов, так как киберполигоны могут быть созданы по образу объектов критической информационной инфраструктуры и больше нацелены на развитие навыков защиты, в то время как CTF предполагает еще и атакующие действия. Киберполигоны зачастую основываются либо на статических шаблонах сетевой инфраструктуры, либо на динамических шаблонах, состоящих из «уязвимых узлов». Здесь же стоит отметить преимущество киберполигонов – доступность обучения. Для тренировки не нужно организовывать мероприятие с двумя командами, а достаточно лишь браузера и нескольких программ (например, для работы с удаленным рабочим столом).

Не менее важным аспектом становится описание и моделирование узлов и сетей, подверженных уязвимостям, а также соответствующих атак. Методам и нотациям формального описания компьютерных атак посвящено немалое количество научных работ [9–14]. В работах [10, 11] объясняется важность выбора правильной методологии для описания угроз информационным системам, в работах [12–14] предлагаются различные подходы к описанию угроз кибербезопасности. Предложенные подходы сравниваются в статье [9], в заключении которой упоминается удобство использования MAL для описания сценариев кибератак.

Реализация уязвимого узла и автоматизированной атаки на него представлена в научной работе [15]. В статье описан механизм исследуемой уязвимости, проведено формальное описание потенциальной компьютерной атаки на узел, а также приведен программный код, позволяющий злоумышленнику провести автоматизированную атаку с помощью модуля из фреймворка Metasploit [16–19]. В работах [16, 17] рассматриваются возможности среды Metasploit Framework, а в работах [18, 19] предлагается использовать фреймворк как средство для автоматизированного пентеста. Таким образом, Metasploit является мощным инструментом и позволяет реализовывать эксплуатации уязвимостей с готовой полезной нагрузкой.

Опубликованные исследования и научные работы позволяют понять, что актуальные уязвимости и соответствующие им атаки возможно описать с помощью существующих методов формального описания, а затем смоделировать автоматизированную атаку на уязвимую систему с помощью фреймворка Metasploit.

Формирование уязвимого узла

В данной работе для создания уязвимого узла используется уязвимость «Adobe ColdFusion Deserialization of Untrusted Data Vulnerability» (CVE-2023-26360), которая была обнаружена 14 марта 2023 года. Adobe опубликовала рекомендации по безопасности с описанием уязвимости, затрагивающей ColdFusion 2021 Update 5 и ColdFusion 2018 Update 15. CVE-2023-26360 – это уязвимость десериализация ненадёжных данных, которая позволяет злоумышленнику удаленно выполнить произвольный код в уязвимой системе. Уязвимость оценивается как критическая, поскольку для её использования не требуется взаимодействие с пользователем [20]. Злоумышленник может использовать эту уязвимость для направления различных полезных нагрузок как на систему в целом, так и на сервер веб-приложения, и делать это без необходимости прохождения проверки аутентификации. Это означает, что атакующий может удалённо выполнить произвольный код на уязвимой системе или сервере без участия или согласия пользователя, а также без необходимости получения прав привилегированного пользователя на сервере [21]. Уязвимости, позволяющие злоумышленникам выполнить код на удаленной системе, считаются одними из наиболее опасных и актуальных в области кибербезопасности.

Суть уязвимости заключается в том, как ColdFusion производит десериализацию ненадежных данных. Для эксплуатации данной уязвимости злоумышленник может отправить на сервер ColdFusion специально сгенерированный запрос, содержащий ненадежные данные, которые в последствии будут десериализованы и выполнены в виде кода. Ход эксплуатации уязвимости зависит от того, какой тип файла используется для обработки. Так, для удаленного выполнения произвольного кода на уязвимой системе злоумышленник может внедрить вредоносные CFML-теги, например, в лог-файл, записи из которого будут преобразованы компилятором NeoTranslator в соответствующие инструкции для выполнения вредоносного кода.

Предварительно, с использованием программного продукта виртуализации VirtualBox, были настроены две виртуальные машины (табл. 1).

Для наглядности конечной целью эксплуатации уязвимости будет являться получение хостом злоумышленника возможности удаленно выполнять код с помощью командной оболочки уязвимого хоста, то есть получение Meterpreter-сессии через TCP-соединение.

Это возможно путем выполнения HTTP-запроса с данными, содержащими произвольные теги CFML. Содержимое этих данных будет записываться в файл журнала ColdFusion. После перевода coldfusion-out.log

Таблица 1

Конфигурация уязвимого узла

Хосты / Компоненты	Злоумышленник	Виртуальная машина с уязвимостью
Операционная система	Kali GNU/ Linux 6.1.0	Ubuntu 22.04.2 LTS
Сетевая конфигурация	inet 10.0.2.11 netmask 255.255.255.0 broadcast 10.0.2.255	inet 10.0.2.15 netmask 255.255.255.0 broadcast 10.0.2.255
Программное обеспечение	msfconsole 6.3.25-dev python 3.11	Adobe ColdFusion 2021 Update 5

файла в формат CFML, выполнятся инструкции, указанные в произвольных тегах CFML, которые после осуществления запроса будут присутствовать в файле журнала.

Для получения Meterpreter-сессии необходимо сгенерировать два HTTP-запроса к серверу. Первый запрос организует точку для подключения на уязвимом узле (удалённо создаст полезную нагрузку и эксплуатирует её, используя оболочку bash). Второй запрос – переведёт файл журнала в формат CFML, что является эксплуатацией уязвимости.

Таким образом, если злоумышленник предварительно запустил хэндлер (программу, ожидающую соединения от жертвы по определённому сокету) – между жертвой и злоумышленником установится вредоносная сессия.

Для оптимизации процесса эксплуатации был выбран такой инструмент для выполнения эксплойтов, как Metasploit[11]. На момент написания статьи в базе знаний Metasploit Framework уже существует модуль «exploit/multi/http/adobe_coldfusion_rce_cve_2023_26360», который содержит в себе код эксплуатации уязвимости на языке Ruby. Для корректной эксплуатации уязвимости необходимо настроить определенные параметры модуля:

1. CFC_ENDPOINT – путь до запрашиваемого целевого CFC-файла.
2. CF_LOGFILE – путь до целевого файла журнала.
3. RHOSTS – адрес уязвимого хоста.
4. RPORT – TCP порт, на котором установлен уязвимый сервер. В рамках представленной эксплуатации подключение к серверу осуществляется по порту 8500.
5. LHOST – адрес локального узла, IP-адрес машины злоумышленника.
6. LPORT – локальный порт. Именно этот порт будет прослушивать злоумышленник, в ожидании установления обратного соединения с жертвой.
7. TARGET – цель для эксплуатации. Выбор цели зависит в основном от целевой ОС и возможностях сессии, получаемой после эксплуатации.

Эксплойт предоставляет возможность получить Meterpreter-сессию через готовый payload, без повышения уровня сессии через shell-соединение. В качестве загружаемой и эксплуатируемой полезной нагрузки был выбран модуль java/meterpreter/reverse_tcp.

В ходе эксплуатации, на машине злоумышленника запускается хэндлер и ждёт входящего подключения по сокету 10.0.2.11:4444. Удалённому серверу будет отправлен POST-запрос к файлу по пути: /cf_scripts/scripts/ajax/ckeditor/plugins/filemanager/iedit.cfc с тэгом method=fxyhppau&_cfclient=true. Это запрос на получение файла iedit.cfc, содержащий некорректные CFML-тэги.

В параметре _variables был передан CFML-код, который создает объект java.net.URL класса Java и инициирует его с заданным адресом. Затем создается массив и класс java.net.URLClassLoader, который загружает класс metasploit.Payload.

Второй запрос преобразует некорректные CFML-тэги в исполняемый код. В результате выполнения этого кода на удаленном сервере эксплуатируется полезная нагрузка «meterpreter/reverse_tcp». Далее, устанавливается связь с TCP-обработчиком на 10.0.2.11:4444. В конечном итоге, между уязвимым сервером и злоумышленником будет установлена Meterpreter-сессия.

Для большей понятности и структурированности при описании эксплуатации уязвимости в данной работе применяется формализация с помощью MAL. Формальное описание способствует более эффективному анализу и обучению, а также упрощает коммуникацию между специалистами разных областей. На рис. 1 представлен мета-граф, который включает в себя следующие элементы [22]:

11 Metasploit Framework [Электронный ресурс]. URL: https://www.metasploit.com/ (дата обращения: 29.09.2023).

1. Активы в системе (большие синие круги): обозначают используемые ресурсы в системе, такие как базы данных компании, рабочая станция администратора и так далее.
2. Логические шаги типа И (красные квадраты): показывают переход между активами и представляют этапы, на которых злоумышленник достигает своих целей, например, несанкционированная регистрация нового пользователя или извлечение информации из базы данных компании.
3. Логические шаги типа ИЛИ (маленькие круги): описывают факторы, благодаря которым атака злоумышленника была успешной, такие как захват прав администратора компьютера или неправильная работа механизма миграции в системе контроля версий Gitea.
4. Шаги защиты (треугольники): отражают возможные меры по предотвращению или противодействию атаке, например, настройка групповой политики безопасности или брандмауэра.
5. Используемые системы (нижние круги оранжевого цвета): указывают на конкретные операционные системы, используемые при успешной атаке, такие как Ubuntu или Astra Linux.

Автоматизированная эксплуатация уязвимости

В рамках выполнения работы стояла задача автоматизации процесса эксплуатации уязвимости. Metasploit Framework состоит из Ruby-скриптов, то есть не поддерживает выполнение Python-скриптов через свою оболочку.

Для решения этой задачи была использована библиотека Pymetasploit. С её помощью можно связываться с Metasploit в программном коде по API с использованием протокола RPC через специализированную службу msfrpcd.

Ниже представлено описание кода для эксплуатации уязвимости и автоматизации действий атакующего на языке Python. В процессе разработки использовался Python 3.11.

Импорт библиотек и установка соединения с msfrpc.

```
import time
from pymetasploit3.msfrpc
import MsfRpcClient, ShellSession, MsfConsole
client = MsfRpcClient 'password', port=55553,
ssl=True)
```

Объявления глобальных переменных. Указанные значения будут использоваться для указания параметров в msfconsole. Если выбранный эксплойт использует корректные параметры по умолчанию – в коде их можно не изменять.

```
RHOSTS = '10.0.2.15'
LHOST = '10.0.2.11'
LPORT = '4444'
```

Объявление и написание функции эксплуатации уязвимости. Здесь происходит организация процесса эксплуатации и настройка параметров msfconsole.

```
def exploit_adobe_cve(config) -> bool:
   exploit = client.modules.use('exploit'
'multi/http/adobe_coldfusion_rce_cve_2023_
26360')
   exploit['RHOSTS'] = 'RHOSTS'
   tries = 5
   pload = client.modules.use('payload', 'java/
meterpreter/reverse_tcp')
   pload['LHOST'] = 'LHOST'
   pload['LPORT'] = 'LPORT'
   for i in range(tries):
      if i >= 1:
         time.sleep(10)
      count = len(client.sessions.list)
```

Рис. 1. – Формальное описание эксплуатации уязвимости с помощью методологии моделирования MAL

```
        exploit.execute(payload=pload)
        if len(client.sessions.list) > count:
            return True
    print('Ошибка эксплуатации')
```

После успешного выполнения скрипта между злоумышленником и уязвимой виртуальной машиной будет установлена Meterpreter-сессия. После этого злоумышленник сможет развивать свой вектор атаки в любом направлении и проводить разнообразные манипуляций как с сервером, так и с самой системой.

Меры защиты

Уязвимости «Adobe ColdFusion Deserialization of Untrusted Data Vulnerability» подвержены системы с функционирующими на них серверами ColdFusion 2021 Update 5 и ColdFusion 2018 Update 15, а также более ранние версии. Одной из мер защиты будет являться установка обновлений безопасности с официального сайта продукта (ColdFusion 2021 Update 6, либо ColdFusion 2018 Update 16 и более поздние версии). В случае, если установка обновлений невозможна или требует времени, можно предпринять самостоятельные меры по усилению безопасности ColdFusion. Одной из таких мер является добавление переменной allowNonCFCDeserialization в код класса JSONUtils.java и добавление проверки на расширение «.cfc» Это реализуемый подход, но требует опыта в разработке и понимания кода ColdFusion.

Второй мерой защиты является отключение компилятора NeoTranslator. Отключение компилятора NeoTranslator не позволит ColdFusion переводить страницы в классы Java.

Предложенные меры защиты относятся к методам защиты от веб-уязвимостей на основе намерений, где под «намерением» подразумевается функциональность, которая должна быть заложена в приложении с учетом целей и решаемых задач [23].

Стоит понимать, что предложенные в данной научной статье меры не являются официальными рекомендациями разработчика и могут ограничивать функциональность ColdFusion. Соответственно, при первой возможности рекомендуется установить обновления от разработчика, которые выпущены специально для устранения критической уязвимости CVE-2023-26360.

Заключение

В результате исследования была детально проанализирована и описана уязвимость «Adobe ColdFusion Deserialization of Untrusted Data Vulnerability» (CVE-2023-26360). Эта уязвимость представляет серьезную угрозу для информационной безопасности, так как позволяет злоумышленникам удаленно выполнять произвольный код на уязвимых серверах ColdFusion. Эксплуатация уязвимости была формально описана с помощью методологии моделирования MAL, что позволит многим специалистам понять последовательность шагов злоумышленника, какие он использовал ресурсы и какие меры противодействия можно предпринять.

Была реализована автоматизированная эксплуатация уязвимости с использованием языка программирования Python и фреймворка Metasploit. Это позволило создать уязвимый узел, который является ценным ресурсом, обеспечивая комплексное и актуальное обучение в киберполигоне. Также практический сценарий атаки может быть использован для оценки уровня защиты системы от подобных угроз. Кроме того, в работе были изложены рекомендации по усилению защиты и описана стратегия для предотвращения атак на Adobe ColdFusion.

Киберполигоны представляют собой перспективное направление в области информационной безопасности, поскольку они играют ключевую роль в повышении уровня подготовки персонала и тестировании на защищенность. Вклад в развитие киберполигонов способствует увеличению компетентных экспертов и приводит к повышению общей безопасности, сокращению успешных кибератак и более надежной защите информационной инфраструктуры.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках базовой части государственного задания ТУСУРа на 2023–2025 гг. (проект № FEWM-2023-0015).

Литература

1. *Карпов, Д. С. Повышение качества подготовки специалистов по направлению подготовки «Информационная безопасность» / Д. С. Карпов, А. А. Микрюков, П. А. Козырев // Открытое образование. – 2019. – Т. 23, № 6. – С. 22–29. – DOI 10.21686/1818-4243-2019-6-22-29. – EDN YEMKVN.*
2. *Harjinder L. et al. Pedagogic Challenges in Teaching Cyber Security--a UK Perspective //arXiv preprint arXiv:2212.06584. – 2022.*
3. *Аверьянов, В. С. Pentest – лаборатория для обучения специалистов направления подготовки информационная безопасность / В. С. Аверьянов, И. Н. Карцан // Актуальные проблемы авиации и космонавтики : Сборник материалов VI Международной научно-практической конференции, посвященной Дню космонавтики. В 3-х томах, Красноярск, 13–17 апреля 2020 года / Под общей редакцией Ю. Ю. Логинова. Том 2. – Красноярск: Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева», 2020. – С. 198–200.*

4. Серпенинов, О. В. Система компетентно-ориентированного обучения специалистов в области информационной безопасности / О. В. Серпенинов // Научный вектор: Сборник научных трудов магистрантов / Под научной редакцией А. У. Альбекова. Том Выпуск 4. – Ростов-на-Дону: Ростовский государственный экономический университет «РИНХ», 2018. – С. 199–202.
5. Меньшенина, С. Г. Структура готовности к профессиональной деятельности специалистов по информационной безопасности / С. Г. Меньшенина // Вестник Самарского государственного технического университета. Серия: Психолого-педагогические науки. – 2018. – № 1(37). – С. 100–107.
6. Ciuperca E., Stanciu A., Cîrnu C. Postmodern education and technological development. Cyber range as a tool for developing cyber security skills //INTED2021 proceedings. – IATED, 2021. – С. 8241–8246.
7. Kornegay M. A., Arafin M. T., Kornegay K. Engaging underrepresented students in cybersecurity using Capture-the-Flag (CTF) competitions (experience) //2021 ASEE Virtual Annual Conference Content Access. – 2021.
8. Karampidis K. et al. Digital Training for Cybersecurity in Industrial Fields via virtual labs and Capture-The-Flag challenges //2023 32nd Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE). – IEEE, 2023. – С. 1–6.
9. Методы формализации описания сценариев кибератак / А. Ю. Якимук, С. А. Устинов, Т. П. Лазарев, А. С. Коваленко // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. – 2022. – № 1-2. – С. 73–76.
10. A Survey on Threat-Modeling Techniques: Protected Objects and Classification of Threats / A. Konev, A. Shelupanov, M. Kataev [et al.] // Symmetry. – 2022. – Vol. 14, No. 3. – DOI 10.3390/sym14030549.
11. Computer network threat modelling / A. Novokhrestov, A. Konev, A. Shelupanov, A. Buymov // Journal of Physics: Conference Series, Tomsk, 20–22 ноября 2019 года. – Tomsk, 2020. – P. 012002. – DOI 10.1088/1742-6596/1488/1/012002.
12. Xiong W. et al. Cyber security threat modeling based on the MITRE Enterprise ATT&CK Matrix //Software and Systems Modeling. – 2022. – Т. 21. – №. 1. – С. 157–177.
13. Johnson P., Lagerström R., Ekstedt M. A meta language for threat modeling and attack simulations //Proceedings of the 13th International Conference on Availability, Reliability and Security. – 2018. – С. 1–8.
14. Xiong W., Lagerström R. Threat modeling–A systematic literature review //Computers & security. – 2019. – Т. 84. – С. 53–69.
15. Уязвимость «Gitea Git Fetch Remote Code Execution»: анализ, формализация автоматизированной эксплуатации, меры защиты / А. А. Конев, А. С. Коваленко, В. С. Репкин, Г. Ю. Семенов // Вестник УрФО. Безопасность в информационной сфере. – 2023. – № 2(48). – С. 67–73. – DOI 10.14529/secur230207.
16. Ромейко, Д. А. Обзор возможностей среды Metasploit Framework / Д. А. Ромейко, Т. И. Паюсова // Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых, Тюмень, 18–23 мая 2022 года / Министерство науки и высшего образования Российской Федерации, Тюменский государственный университет, Институт математики и компьютерных наук. Том Выпуск 20. – Тюмень: ТюмГУ-Press, 2022. – С. 318–325.
17. Khera Y. et al. Analysis and impact of vulnerability assessment and penetration testing //2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). – IEEE, 2019. – С. 525–530.
18. Valea O., Oprişa C. Towards pentesting automation using the metasploit framework //2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP). – IEEE, 2020. – С. 171–178.
19. Raj S., Walia N. K. A study on metasploit framework: A pen-testing tool //2020 International Conference on Computational Performance Evaluation (ComPE). – IEEE, 2020. – С. 296–302.
20. Li Y., Liu Q. A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments //Energy Reports. – 2021. – Т. 7. – С. 8176–8186.
21. Biswas S. et al. A study on remote code execution vulnerability in web applications //International Conference on Cyber Security and Computer Science (ICONCS 2018). – 2018. – С. 50–57.
22. Wideł W., Mukherjee P., Ekstedt M. Security Countermeasures Selection Using the Meta Attack Language and Probabilistic Attack Graphs //IEEE Access. – 2022. – Т. 10. – С. 89645–89662.
23. Методы защиты веб-приложений от злоумышленников / В. Е. Боровков, П. Г. Ключарев, // Вопросы кибербезопасности – 2023. – № 5(57). – С. 89–99. – DOI 10.21681/2311-3456-2023-5-89-99.

АНАЛИЗ НЕКРИПТОГРАФИЧЕСКИХ МЕТОДОВ ЗАЩИТЫ ИНФОРМАЦИИ В РАДИОКАНАЛАХ ИНФОРМАЦИОННЫХ СИСТЕМ

Махов Д. С.[1],

DOI: 10.21681/2311-3456-2024-1-82-88

Цель исследования *состоит в анализе проблемных вопросов определения понятия некриптографических методов, показателей и критериев защиты информации в радиоканале, а также в кратком анализе существующих методов борьбы с помехами и возможности их использования в качестве некриптографических методов защиты информации в радиоканале.*

Методы исследования: *в работе применен дедуктивный подход к определению понятия «защита информации в радиоканале». Затем на основе логического вывода и индуктивного подхода проведено соотнесение показателей защищенности информации и показателей, применяемых для оценки радиотехнических систем при воздействии помех.*

Результат исследования: *на основе сочетания метода аналогии и дедуктивного подхода установлены взаимосвязи между показателями защиты информации и показателями оценки радиотехнических систем при воздействии помех. Изложен проблемный вопрос о нормативном определении понятия «некриптографических» методов защиты информации в радиоканале. На основе анализа научных публикаций по теме исследования приведено краткое описание методов борьбы с помехами и их влияния на защищенность радиотехнической системы, как информационной. Предложено в качестве математического аппарата оценивания использовать аппарат теории вероятностей. Намечены пути установления аналитической взаимосвязи показателей защищенности информации в радиоканале и параметров радиотехнических систем.*

Практическая ценность: *предложен подход к аналитическому описанию защищенности информации в радиоканалах. Это позволит учитывать показатели как криптографических, так и некриптографических методов защиты информации при анализе защищенности информационных систем. Определено направление научных исследований, которое позволит дать нормативное определение и сформировать классификацию некриптографических методов защиты информации в радиоканалах, что может быть использовано при синтезе систем и средств защиты информации.*

Ключевые слова: *оценочно-критериальная база, радиотехническая система, защищенность информации, конфиденциальность, доступность, помехоустойчивость, скрытность, воздействие помех, пространственная селекция, фильтрация, расширение спектра.*

ANALYSIS OF NON-CRYPTOGRAPHIC INFORMATION PROTECTION METHODS IN WIRELESS INFORMATION SYSTEMS

Makhov D.S.[2]

The purpose *of the research is to analyze the problematic issues of the non-cryptographic methods concept defining, indicators and criteria for information protection in the radio channel, as well as an existing interference resistance methods analysis, and the using possibility of their as non-cryptographic information secure methods in the radio channel.*

1 Махов Денис Сергеевич, доктор технических наук, начальник кафедры защиты информации в радиолиниях систем и комплексов вооружения, военной и специальной техники Краснодарского высшего военного орденов Жукова и Октябрьской Революции Краснознаменного училища имени генерала армии С. М. Штеменко, г. Краснодар, Россия. E-mail: sinedvoham@yandex.ru

2 Denis S. Makhov, Dr.Sc. (in Engineering sciences), Head of department information secure in radio channel of the military equipment systems Krasnodar Higher Military Orders of Zhukov and the October Revolution of the Red Banner School named after General of the Army S.M. Shtemenko, Krasnodar, Russia. E-mail: sinedvoham@yandex.ru

Research methods: *the article uses a deductive approach to the definition of the «radio channel information secure» concept. Then, on the logical inference basis and an inductive approach, the correlation of information security indicators and radio engineering systems evaluation indicators under the interference influence was carried out.*

The research result: *on the analogy method and the deductive approach combination basis, the interrelationships between information security indicators and radio engineering systems evaluation indicators under the interference influence have been established. The normative definition problem of the information security «non-cryptographic» methods concept in the radio channel is presented. Based on the research field scientific publications review, the interference resistance methods description and their influence on the security of the radio engineering system as an information system is given. It is proposed to use the probability theory methods as a mathematical evaluation instrument. The ways of establishing an analytical relationship between the information security indicators in the radio channel and the radio engineering systems parameters are defined.*

Practical significance: *an approach to the information security analytical description in radio channels is proposed. This way allows taking into account the indicators of both cryptographic and non-cryptographic information protection methods during the information systems security analyze. The scientific research direction to give a normative definition and form a non-cryptographic information security methods classification in radio channels has been determined. This can be used in the information security systems and tools synthesis.*

Keywords: *evaluation and criteria base, radio engineering system, information security, confidentiality, accessibility, interference immunity, stealth, interference influence, space selection, filtering, spectrum spreading.*

Введение

В настоящее время увеличивается количество информационных систем как гражданского, так и военного назначения, осуществляющих информационный обмен по радиоканалам (РК), вследствие чего возникает множество проблем обеспечения информационной безопасности таких систем [1, 2]. Это повышает актуальность вопроса защиты всех видов информации, передаваемой по РК, от воздействий внешних и внутренних вредных факторов. Нормативными документами предписано делить методы защиты информации в РК на криптографические и некриптографические. И если криптографические методы защиты информации в РК определены и известны [3], то понятие некриптографических методов для области защиты информации в РК не определено. Возникают проблемные вопросы определения понятия «некриптографические методы», классификации самих методов в области защиты информации в РК, решения практических задач на их основе. Существует некоторое количество публикаций, сводящих вопросы защиты информации в РК не криптографическими методами к области технической защиты от побочных электромагнитных излучений [4] или к вопросам защиты от несанкционированного доступа [5].

Как правило, показатели и критерии оценивания защищенности информации в РК в данных публикациях весьма расплывчаты, что затрудняет формирования критической оценки преимущества тех или иных «некриптографических» методов[3] [6]. Следует отметить, что в случае, когда средой распространения информации между двумя элементами информационной системы является воздушное пространство, то такая система коренным образом отличается от других типов систем не только по оценочно-критериальной базе ее функций, но и по инструментарию обеспечения ее эффективного функционирования [7]. Это связано с тем, что основным внешним вредным фактором для такой системы являются электромагнитные помехи на основе известного факта, что в свободном пространстве информация переносится с помощью электромагнитных волн. Как известно [8, 9], воздействие помех на функционирование информационной системы в РК осуществляется, как правило, в свободном пространстве на сигнальном, или, согласно понятийному аппарату информационных систем, на физическом уровне эталонной модели взаимодействия открытых систем (ЭМВОС).

Следует также упомянуть, что информационные системы, информация в которых циркулирует по РК, называются радиотехническими системами (РТС). Оценке эффективности функционирования РТС посвящено достаточное количество литературы, отражающей более ста лет накопленных знаний в области радиотехники. Определены показатели, критерии оценки, стандарты и рекомендации, разработаны методы достижения значений показателей, определены оптимальные значения (например, порог Шеннона) а также пути научного развития. Вместе с тем, анализ показал, что такие показатели для вопросов защиты информации в РТС отсутствуют, а показатели криптографических методов не учитывают особенности функционирования РТС.

3 Брауде-Золотарев Ю.М. Алгоритмы безопасности радиоканалов // Алгоритм безопасности. 2013. № 1. С. 64–66.

На основании вышеизложенного возникает противоречие, заключающееся в том, что с одной стороны в РТС строго определены информационные показатели для криптографических методов защиты информации в РК, которые не учитывают влияние внешних вредных факторов физического и канального уровней ЭМВОС, а с другой стороны внешние вредные факторы учитываются в радиотехнических показателях оценки РТС, но не определены для информационных показателей некриптографических методов защиты информации в РК.

Анализ оценочно-критериальной базы защиты информации в радиоканалах

Для разрешения указанного противоречия необходим анализ понятий и определений из области защиты информации, которые служат или могут служить показателями оценки систем передачи информации в РК.

Как известно объектом защиты информации являются носители информации, под которыми понимается в том числе физическое поле, отражающее информацию в виде символов, образов и сигналов[4].

На основе этого тривиален вывод, что любая система связи есть система передачи информации. Это также подтверждается документально. В частности, из нормативных документов следует, что радиосвязь – электросвязь, осуществляемая посредством радиоволн[5]. А в нормативном документе по электросвязи приводится определение электросвязи[6], такое что «Электросвязь – это любые излучения, передача или прием знаков, сигналов, голосовой информации, письменного текста, изображений, звуков или сообщений любого рода по радиосистеме, проводной, оптической и другим электромагнитным системам». При этом, как известно, информационная система – это система, в том числе обрабатывающая и выдающая информацию для дальнейшего использования. Отсюда следует, что РТС является информационной системой. В РТС информация отражена битовым потоком, в двустороннем порядке модулирующим несущие колебания, преобразуемые в дальнейшем в электромагнитные волны. Так что все информационные показатели имеют место и могут преломляться для РТС, в том числе касаемо функций и методов защиты информации в РК.

Таким образом, связь – есть передача информации за исключением случаев, когда канал связи имеется, а информация по нему не передается.

Теперь рассмотрим понятия области защиты информации касательно методов защиты информации в РК. В нормативных документах по защите информации указано, что защита информации – деятельность, направленная на предотвращение утечки защищаемой информации, несанкционированных (но преднамеренных) и непреднамеренных воздействий на нее.

Для обеспечения безопасности информации, под которой можно понимать конечный результат, достижение цели защиты, перевод информационной системы (РТС) в состояние защищенности, используются два типа защиты – техническая и криптографическая. Также существует определение криптографической защиты информации, как ЗИ с помощью ее криптографического преобразования.

При этом в определении технической защиты информации фигурирует понятие некриптографических методов. А именно, под технической защитой информации, как ЗИ заключающейся в обеспечении некриптографическими методами безопасности информации, подлежащей защите с применением технических, программных и программно-аппаратных средств. Но вместе с тем определение некриптографических методов, как всех, которые не подпадают под определение криптографических, весьма расплывчато и не определено (рис. 1).

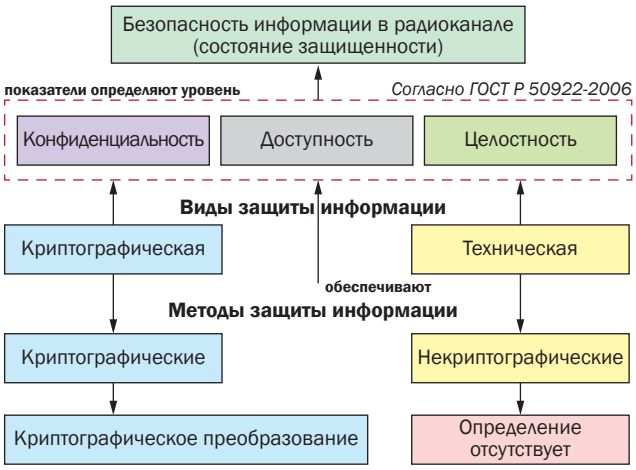


Рис.1. К определениям методов защиты информации, отраженных в нормативных документах

На основе вышеизложенного к некриптографическим можно отнести традиционные и вновь создаваемые методы повышения уровня таких показателей, как помехозащищенность, помехоустойчивость и скрытность [10]. Однако возникает неопределенность оценивания эффективности таких методов с точки зрения защиты информации, обусловленная отсутствием теоретического базиса и аналитического описания показателей защиты информации в РК для РТС.

В силу того, что показатели защиты информации имеют лишь нормативное определение, возможна попытка их аналитического описания с помощью вероятностного подхода. Обозначая использование

4 ГОСТ Р 50922–2006. Защита информации. Термины и определения. – М.: Стандартинформ, 2008. – 12с.

5 ГОСТ 24375-80. Радиосвязь. Термины и определения. – М.: Издательство стандартов, 1987. – 57 с.

6 ГОСТ Р 53111-2008. Устойчивость функционирования сети электросвязи. Требования и методы проверки. – М.: Стандартинформ, 2011. – 31 с.

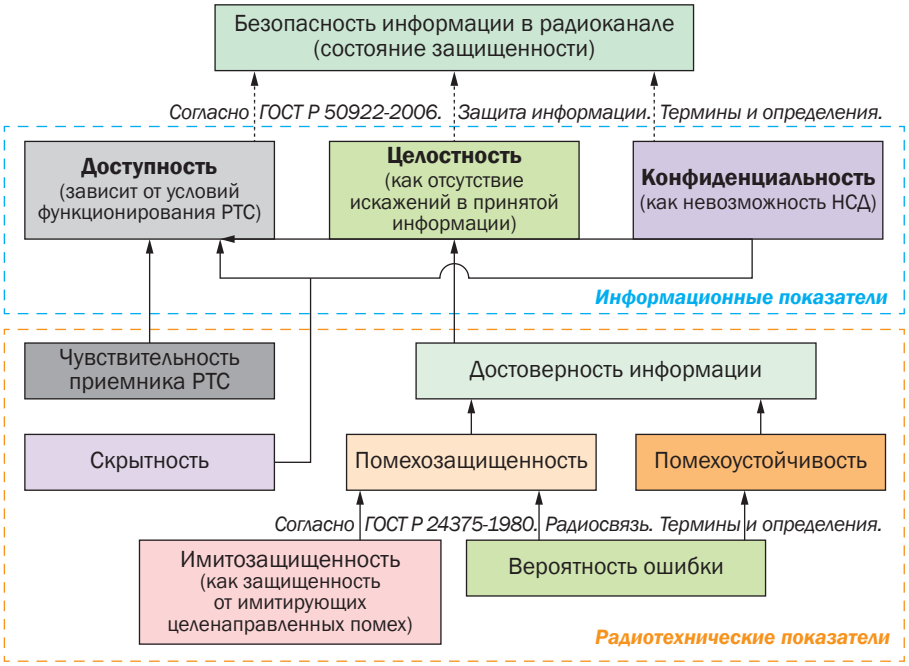


Рис.2. Вариант установления связей между показателями, характеризующими свойства информации в РК и показателями, характеризующими РТС

для защиты РК криптографических методов в виде M, а некриптографических – N, функцию, описывающая состояние защищенности информации можно представить в следующем виде:

$$Z = \rho(M) + \rho(N) - \rho(MN), \quad (1)$$

где: ρ – вероятность.

Описание в виде (1) может правомерно служить основой для расчета защищенности информации в РК. Однако для установления связи между параметрами РТС и показателями защищенности информации необходимо рассмотреть основные показатели защиты информации (рис. 2).

Состояние защищенности информации, ее безопасность, обеспечивается ее конфиденциальностью C, доступностью D и целостностью S и является их функцией2:

$$Z = f(D,C,S). \quad (2)$$

Конфиденциальность – обязательное для выполнения лицом, получившим доступ к информации требование не передавать такую информацию третьим лицам без согласия ее обладателя. То есть «конфиденциальность потока сообщений означает, что никто, даже при наличии доступа к каналам передачи и узлам коммутации сети, не должен иметь возможности установить, какого типа и какие данные передаются пользователю или поступают от пользователя, а также объем пересылаемых данных и адреса назначения»[7].

Преломляя понятие конфиденциальности к информации, носителем которой является сигнал в РК, и рассматривая ее в отсутствии криптографических преобразований только с технической стороны, можно сказать следующее. Информация при данных условиях будет обладать таким свойством вне РТС. То есть получить информацию может любой абонент на приемной стороне, параметры РТС которого совпадают с параметрами передающей РТС, и находящийся в зоне действия передающей РТС. При этом согласие абонента передающей РТС для доступа к передаваемой ею информации не требуется. Следовательно, понятие конфиденциальности информации в РК применимо и коррелирует с понятием скрытности или разведзащищенности:

$$C = 1 - \rho_r = 1 - \rho_{ob}\,\rho_{st}\,\rho_{in}, \quad (3)$$

где: ρ_r – вероятность разведки параметров РТС, ρ_{ob} – вероятность обнаружения сигналов РТС, ρ_{st} – вероятность раскрытия структуры сигнала, ρ_{in} – вероятность раскрытия информации (смысла).

Доступность – состояние информации, при котором субъекты, имеющие права доступа, могут реализовать их беспрепятственно.

Доступность информации в РК является условием для конфиденциальности. Если информация будет не доступна, то вопрос о конфиденциальности снимается. На первый взгляд доступность определяется предельной чувствительностью приемника и достаточным уровнем сигнала на его входе на фоне шумов и помех, то есть уровнем мощности принятого сигнала, при котором обеспечивается отношение

7 Воробьев Е. Г. Управляемая поляризация электромагнитных волн как средство повышения скрытности передачи информации // Информатика, управление и компьютерные технологии. Известия СПбГЭТУ «ЛЭТИ». 2014. № 9. С 44–49.

его к уровню шума, равное единице. Тогда доступность можно определить вероятностью наступления такого события:

$$D = \rho(Q = 1), \quad (4)$$

где: Q – отношение сигнал/шум на входе приемника.

Тот факт, что при условии уверенного приема сигнала РТС (выполнения условия доступности) информация может быть искажена, относится к понятию целостности информации.

Целостность – состояние информации, при котором отсутствует любое ее изменение либо изменение осуществляется только преднамеренно субъектами, имеющими на это право. В данном контексте понятие целостности схоже с понятием достоверности и определяется вероятностью искажения бита или символа:

$$S = 1 - \rho_{b(s)}, \quad (5)$$

где: $\rho_{b(s)}$ – вероятность ошибки на бит (символ).

Для того, чтобы увязать формулы (3)–(5), можно использовать формулу полной вероятности, аналогичную (1).

Разумеется, что данные метрики являются лишь предложением и могут быть усложнены (например, в виде представления конфиденциальности условной вероятностью). Логический вывод и описание информационных показателей защищенности информации в РК приведено с целью разрешения указанного во введении противоречия. Возможные пути решения в виде аналитического описания конфиденциальности, доступности и целостности информации в РК позволят установить взаимосвязь между защищенностью информации и параметрами различных уровней ЭМВОС РТС. Также это позволит нормативно определить некриптографические методы защиты информации в РК, проводить анализ таких методов и разрабатывать новые.

Анализ современных «некриптографических» методов защиты информации в радиоканалах

В настоящее время, учитывая указанную выше проблематику, к некриптографическим методам защиты информации в РК можно отнести классические методы обеспечения помехоустойчивости, имитоустойчивости и скрытности.

Имитоустойчивость, определяемая как защита от подмены или ввода ложной информации, может быть рассмотрена как защита от имитирующих помех[8]. Имитоустойчивости также посвящены работы [11, 12]. В [10] также подробно описаны «некриптографические» методы защиты информации в виде борьбы с помехами на различных уровнях приемной РТС.

8 Максимов М. В. Защита от радиопомех: под ред. М. И. Максимова. М.: Сов. радио, 1976. 496 с.

Разделим данные методы на антенные методы, методы фильтрации, методы помехоустойчивого кодирования и сигнальные методы.

К антенным методам относятся методы компенсации помех, пространственной фильтрации и селекции, частотной селекции и поляризационной селекции.

Компенсация помех осуществляется за счет формирования провалов в диаграмме направленности в направлении прихода помехового сигнала.

Пространственная селекция осуществляется за счет формирования главных максимумов диаграммы направленности в направлении цели или источника сигнала, управления главными максимумами диаграммы направленности и управлением уровнями боковых лепестков [12, 13]. Эти функции основаны на методах решения задачи синтеза антенн и управления весовыми коэффициентами – комплексными амплитудами антенных решеток. Последняя функция АР базируется на основе теории адаптивных АР [13, 14].

Поляризационная селекция позволяет использовать управляемые поляризационные характеристики антенн для повышения качества приема сигналов [8, 10, 14]. Если антенны приемника и передатчика имеют одинаковую поляризацию, мощность принятого сигнала будет максимальной при прочих равных. Известно направление поляризационной модуляции сигнала, результаты которого могут быть использованы для обеспечения защищенности информации в РК. Данные методы позволяют использовать параметры поляризационного эллипса для кодирования информации. Сами параметры изменяют с помощью адаптивного процессора при условии выполнения антенны с круговой поляризацией.

Частотная селекция позволяет использовать антенну как частотный фильтр при соблюдении условия широкополосности. Необходимо обеспечить максимальную амплитуду сигнала на входе антенны при сканировании по частоте. Для этого необходимо учитывать амплитудно-частотную характеристику антенно-фидерного устройства.

Кратким выводом по антенным методам является заключение о том, что данные методы разработаны относительно недавно, в середине прошлого века, и по аналогии с цифровой связью сегодня находят физическую основу для реализации. Такой основой выступают цифровые антенные решетки и смарт-антенны [15], включающие в состав микроконтроллеры управления параметрами, схемы цифрового управления лучом и специальные вычислители на основе методов искусственного интеллекта. Результатом применения совокупности методов и технологий является высокий уровень мощности сигнала на входе приемника, дополнительное кодирование

информации параметрами антенн, что позволяет на физическом уровне внести вклад в защиту информации в РК.

Методы фильтрации основаны на оптимизации и разработке алгоритмов управления фильтрами радиоприемных устройств. Суть методов состоит в приближении амплитудно-частотной характеристики фильтра к идеальной. Это позволяет повысить качество выделения полезного сигнала из сигнальной смеси, и в результате уменьшить количество ошибок на входе канального кодера. В качестве наиболее перспективных фильтров можно выделить адаптивный фильтр Калмана. Параметры фильтров, такие как крутизна и неравномерность АЧХ, частота среза АЧХ и порядок фильтра, оказывают косвенное влияние на достоверность принятой информации.

Методы помехоустойчивого кодирования дают вклад в защиту информации в РТС на канальном уровне ЭМВОС. Разнообразие данных методов и их модификаций достаточно велико и известно[9]. Следует заметить, что уменьшение исправление ошибок декодером можно связать с целостностью информации. Наиболее применяемыми в РК являются каскадные коды, состоящие из кодов, корректирующих одиночные ошибки и кодов, корректирующих групповые ошибки. Первые представлены кодами с низкой плотностью единиц в порождающей матрице (LDPC), а вторые представлены кодом Рида-Соломона и его модификациями.

Класс сигнальных методов можно отобразить методами изменения параметров сигнала на основе применения различных типов модуляции [16, 17], их модификаций, а также методов расширения спектра[10] [16]. Синтез широкополосных сигналов при разумных технических затратах на их реализацию позволяет осуществлять передачу информации на энергетическом уровне предела Шеннона, обеспечивая требуемую скрытность, что влияет на конфиденциальность информации в РК. Использование псевдослучайной перестройки рабочей частоты (ППРЧ) также позволяет обеспечить целостность и конфиденциальность информации в РК за счет скачков по поднесущим частотам и восстановлении информации в случае воздействия помехи в узкой полосе. Кроме того, традиционным подходом повышения скрытности передаваемой информации является скремблирование.

Необходимо отметить совмещение нескольких различных методов для достижения эмерджентных свойств РТС по защищенности информации в РК. Так, применение методов модуляции при разделении по ортогональным поднесущим частотам (OFDM – orthogonal frequency division modulation) в совокупности с методами на основе технологии MIMO (Multiple Input – Multiple Output) позволяет использовать в сочетании частотный и пространственный ресурс РК и реализовывать множество способов передачи, совмещая методы пространственного и частотного разнесения для повышения скрытности[11] [18]. На основе указанных методов в 2022 году утвержден стандарт передачи информации Wi-Fi 6.

Выводы

Таким образом, в работе проведен анализ нормативных документов по защите информации и радиосвязи для определения показателей защищенности информации в РК при использовании некриптографических методов ее защиты. Определено, что такие «информационные» показатели могут быть на основе логического вывода интерпретированы в показатели РТС, такие как скрытность, помехоустойчивость, имитостойкость и помехозащищенность. Либо может быть установлена аналитическая взаимосвязь между «информационными» и радиотехническими показателями. Так же на основании того, что определение «некриптографических» методов защиты информации в РК не дано ни в одном нормативном документе, то во второй части работы проведен неполный анализ радиотехнических методов, способных выступить в качестве таковых. Следует отметить, что методы оптимального приема, оптимальной фильтрации, оптимального кодирования, теории синтеза антенн и другие достаточно хорошо разработаны и описаны в научной литературе. Однако отличительной особенностью сегодняшнего научного развития является обострение междисциплинарного подхода, когда при проектировании лавинообразно увеличивающегося количества и видов радиотехнических систем невозможно четко разграничить защиту информации, радиотехнику, теорию автоматического управления, программирование микроконтроллеров, теорию цепей и технологии искусственного интеллекта. В связи с этим разработка новых методов защиты информации в РК на основе совмещения известных или заимствованных методов из других областей научного знания позволит внести значительный вклад в информационную безопасность.

9 Касами Т., Токура Н., Ивадари Ё., Инагаки Я. Теория кодирования: пер. с яп. М.: Мир, 1978. 568 с.

10 Борисов В. И. Помехозащищенность систем радиосвязи с расширением спектра сигналов методом псевдослучайной перестройки рабочей частоты. М.: Радио и связь, 2000. 384 с.

11 Андронов, И. С., Финк Л. М. Передача дискретных сообщений по параллельным каналам. М. : Советское радио, 1971. 408 с.

Литература

1. *Мариненков Е. Д., Викснин И. И., Жукова Ю. А., Усова М. А. Анализ защищенности информационного взаимодействия группы беспилотных летательных аппаратов // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 5. С. 817–825. DOI: 10.17586/2226-1494-2018-18-5-817-825.*
2. *Головской В. А., Филинов В. С. Предложения по созданию когнитивных систем передачи данных для робототехнических комплексов // T-Comm: Телекоммуникации и транспорт. 2019. Т. 13. №9. С. 22–29.*
3. *Андреев А. М., Мальцев Г. Н., Федоренко М. Ю. Алгоритмы и аппаратура криптографической защиты информации в командных и телеметрических радиолиниях зарубежных космических систем // Успехи современной радиоэлектроники. 2018. № 4. С. 14–26.*
4. *Швиденко С. А., Иванов С. В., Хорольский Е. М., Савельев И. В. Один из эффективных подходов к защите информации в радиолиниях робототехнических комплексов с группами беспилотных летательных аппаратов на основе блокчейн технологии // Информатика, вычислительная техника и управление. 2022. Т.14. № 5. С. 21–26.*
5. *Коротков В. В., Мельников А. В. Актуальные вопросы информационной безопасности радиосвязи морского и речного транспорта // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и Технические Науки. 2021. №12. С. 82–84. DOI: 10.37882/2223-2966.2021.12.14 2 (45).*
6. *Макаренко С. И. Информационный конфликт системы связи с системой дестабилизирующих воздействий. Часть. I: Концептуальная модель конфликта с учетом ведения разведки, физического, радиоэлектронного и информационного поражения средств связи // Техника радиосвязи. 2020. № 45. С. 104–117. DOI: 10.33286/2075-8693-2020-45-104-117.*
7. *Иванов М. А. Способ обеспечения универсальной защиты информации, пересылаемой по каналу связи // Вопросы кибербезопасности. 2019. № 3 (31). С. 45–50.*
8. *Макаренко С. И. Анализ средств и способов противодействия беспилотным летательным аппаратам. Часть 3. Радиоэлектронное подавление систем навигации и радиосвязи // Системы управления, связи и безопасности. 2020. № 2. С. 101–175. DOI: 10.24411/2410-9916-2020-10205.*
9. *Ватрухин Е. М. Комплексная защита информации в каналах «земля-борт» // Вестник Концерна ВКО «Алмаз-Антей». 2020. № 4. С. 6–14. DOI: 10.38013/2542-0542-2020-4-6-14.*
10. *Богатырев А. А., Ермолаев А. С., Саменков Е. В., Нуржанов Д. Х., Подсякина А. Ю. Физические принципы методов защиты от помех // Труды Международного симпозиума «Надежность и качество». 2018. Т. 2. С. 315–317.*
11. *Глобин Ю. О., Финько О. А. Способ обеспечения имитоустойчивой передачи информации по каналам связи // Наукоемкие технологии в космических исследованиях Земли. 2020. Т. 12. № 2. С. 30–43. DOI: 10.36724/2409-5419-2020-12-2-30-43.*
12. *Басан Е. С., Прошкин Н. А., Силин О. И. Повышение защищенности беспроводных каналов связи для беспилотных летательных аппаратов за счет создания ложных информационных полей // Сибирский аэрокосмический журнал. 2022. Т. 23, № 4. С. 657–670. DOI: 10.31772/2712-8970-2022-23-4-657-670.*
13. *Шмачилин П. А., Шумилов Т. Ю. Матричная диаграммообразующая схема цифровой антенной решётки // Труды МАИ. 2019. № 109. DOI: 10.34759/trd-2019-109-12*
14. *Ваганова А. А., Кисель Н. Н., Панычев А. И. Направленные и поляризационные свойства микрополосковой реконфигурируемой антенны, перестраиваемой по частоте и поляризации // Известия ЮФУ. Технические науки. 2021. № 2. С. 74–83. DOI: 10.18522/2311-3103-2021-2-74-83*
15. *Ma Y., Wang J., Li Y., Chen M., Li Z., Zhang Z. Smart antenna with automatic beam switching for mobile communication // EURASIP Journal on Wireless Communications and Networking. 2020. No. 179. Pp. 2–4. DOI: 10.1186/s13638-020-01792-4*
16. *Карпухин Е. О., Макаренков Н. С. Применение сигналов OCDM-OFDM с псевдослучайной перестройкой рабочей частоты для предотвращения атак на физическом уровне // Труды МАИ. 2019. № 106.*
17. *Khalifa M. A. E., Emam A. E., Youssef M. I. Performance enhancement of MIMO-OFDM using redundant residue number system // Advances in science, Technology and engineering systems journal. 2018. Vol. 3, No. 4. Pp. 1–7.*
18. *Elghany M. A., Emam A. E., Youssef M. I. ICI and PAPR enhancement in MIMO-OFDM system using RNS coding // International Journal of Electrical and Computer Engineering (IJECE). 2019. Vol. 9. No. 2. pp. 1209–1219. DOI: 10.11591/ijece.v9i2.pp1209-1219.*

АНАЛИЗ ПРЕДЕЛЬНЫХ ВОЗМОЖНОСТЕЙ МЕТОДОВ ШУМОПОНИЖЕНИЯ И РЕКОНСТРУКЦИИ РЕЧЕВЫХ СИГНАЛОВ, МАСКИРУЕМЫХ РАЗЛИЧНЫМИ ТИПАМИ ПОМЕХ

Хорев А. А.[1], Дворянкин С. В.[2], Козлачков С. Б.[3], Василевская Н. В.[4]

DOI: 10.21681/2311-3456-2024-1-89-100

Цель исследования: *оценка границ применимости методов шумопонижения и реконструкции (далее – методов шумоочистки) речевых сигналов.*

Метод исследований: *артикуляционные испытания.*

Результат и практическая ценность: *авторами на основе теоретических и экспериментальных исследований проведена оценка возможностей улучшения качества речевых сигналов путем применения различных методов шумоочистки и определены границы применимости данных методов. В результате экспериментальных исследований установлено, что все современные методы шумоочистки имеют недостаточную эффективность в случае корректного применения для маскирования фонограммы речеподобной помехи. Кроме того, в ходе исследований определено граничное значение отношения «сигнал/шум», при котором методы шумоочистки становятся неэффективными.*

Вклад авторов: *Хорев А. А. провел артикуляционные испытания и произвел статистическую обработку полученных результатов. Дворянкин С. В. провел испытания средства шумопонижения «Лазурь». Козлачков С. Б. подготовил артикуляционные тексты и провел испытания средства шумопонижения «GritTec's Noise Cancellation». Василевская Н. В. подготовила аналитический обзор и провела испытания средства шумопонижения «Sound Cleaner».*

Ключевые слова: *акустическая речевая разведка, разборчивость речи, речевой сигнал, защита информации, шумоочистка, спектральное вычитание, фильтрация, линейное предсказание.*

THE ANALYSIS OF THE POTENTIAL CAPABILITIES OF METHODS OF NOISE REDUCTION AND RECONSTRUCTION OF ACOUSTIC SPEECH SIGNALS MASKED BY VARIOUS TYPES OF NOISE

Horev A. A.[5], Dvoryankin S. V.[6], Kozlachkov S. B.[7], Vasilevskaya N. V.[8]

Purpose of the study: *the analysis of the potential capabilities of improving the quality of speech signals by applying various methods of noise reduction and reconstruction of acoustic speech signals.*

Research method: *articulation tests.*

1 Хорев Анатолий Анатольевич, доктор технических наук, профессор, Национальный исследовательский университет «МИЭТ», Зеленоград, Россия. E mail: horev@miee.ru, https://orcid.org/0000-0001-9074-385X

2 Дворянкин Сергей Владимирович, доктор технических наук, профессор, Национальный исследовательский ядерный университет «МИФИ», Москва, Россия. E mail: svdvoryankin@mephi.ru, https://orcid.org/0000-0001-6908-0676

3 Козлачков Сергей Борисович, кандидат технических наук, Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия. E mail: ksb.perovo@mail.ru, https://orcid.org/0000-0002-7096-6711

4 Василевская Надежда Валерьевна, Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия. E mail: infuzoriavalenoc@yandex.ru, https://orcid.org/0000-0002-0078-8665

5 Anatoly A. Horev, Dr.Sc., Professor, National Research University of Electronic Technology, Moscow Zelenograd, Russia,. E mail: horev@miee.ru, https://orcid.org/0000-0001-9074-385X

6 Sergey V. Dvoryankin, Dr.Sc., Professor, National Research Nuclear University MEPhI, Moscow, 115409, Russia, e-mail: svdvoryankin@mephi.ru, https://orcid.org/0000-0001-6908-0676

7 Sergey B. Kozlachkov, Ph.D. in Technology, Bauman Moscow State Technical University, Moscow, Russia. E mail: ksb.perovo@mail.ru, https://orcid.org/0000-0002-7096-6711

8 Nadezhda V. Vasilevskaya, Bauman Moscow State Technical University, Moscow. Russia. E mail: infuzoriavalenoc@yandex.ru, https://orcid.org/0000-0002-0078-8665

Result and Practical value: *While making a theoretical and experimental studies the authors evaluated the possibilities of improving the quality of speech signals by applying various methods of noise reduction and reconstruction of acoustic speech signals and determined the limits of their applicability. As a result of experimental studies, it has been established that all modern noise reduction technologies are insufficiently effective when a phonogram is protected by speech-like noise. The research also determined the limiting value of the signal-to-noise ratio at which noise reduction methods become ineffective.*

Authors' contribution: *Horev A. A. conducted articulation tests and performed statistical processing of the results. Dvoryankin S. V. conducted tests of the Lazur noise reduction product. Kozlachkov S. B. prepared articulation texts and tested the noise reduction product «GritTec's Noise Cancellation». Vasilevskaya N. V. prepared an analytical review and tested the noise reduction product «Sound Cleaner».*

Keywords: *acoustic speech intelligence, intelligibility of speech, speech signal, information protection, noise reduction, spectral subtraction, filtration, linear prediction.*

Введение

В настоящее время широкое распространение получили методы шумопонижения и реконструкции акустических речевых сигналов (РС), принятых в условиях помех, которые применяются в различных системах распознавания и обработки речи бытового назначения и, очевидно, используются злоумышленником при перехвате речевых сигналов. К сожалению, границы применимости и предельные достижимые показатели улучшения качества акустических РС для подавляющего большинства методов и алгоритмов шумопонижения (далее – методов шумоочистки) их разработчиками не указываются, в связи с чем провести оценку потенциала злоумышленника не представляется возможным. При этом необходимо учитывать, что оценка предельных возможностей и ограничений методов шумоочистки является ключевым фактором задания и обоснования использования значений показателей защищенности акустической речевой информации.

Ввиду отсутствия достоверных сведений о предельных возможностях методов шумоочистки, используемых злоумышленником, представляется трудновыполнимой задача разработки методики оценки защищенности акустической речевой информации, и, следовательно, создания системы защиты помещений от утечки акустической речевой информации по техническим каналам, обеспечивающей требуемую эффективность защиты от всех возможных видов помех.

Авторами настоящей работы предпринята попытка провести анализ предельных возможностей применяемых в настоящее время методов шумоочистки, определить границы их использования и обосновать наиболее эффективный тип помехи, которая была бы достаточно устойчива по отношению к подавляющему большинству современных методов шумоочистки.

Основные применяемые в настоящее время функциональные классы и методы шумоочистки представлены на рисунке 1.

Рис. 1. Классы методов шумопонижения и реконструкции речи

Методы классов фильтрации, компенсации и коррекции на рис. 1 условно можно отнести к методам шумопонижения, а методы четвертого класса – к методам реконструкции РС по его остаточным следам (в частности трекам речевых вокализмов), определяемым на зашумленных спектрограммах.

Рассмотрим некоторые из этих методов подробнее с учетом их вклада в улучшение отношения сигнал/шум при использовании процедур шумоочистки.

Методы, основанные на оценке спектральных характеристик помех (спектральное вычитание)

Поскольку человеческий слух крайне слабо чувствителен к фазе акустического сигнала, методы спектрального вычитания направлены исключительно на восстановление амплитуды спектра исходного РС. При этом корректировка амплитуды спектра очищенного РС осуществляется с помощью вычитания средней амплитуды шума из мгновенного спектра амплитуды зашумленного сегмента (кадра) РС[9]:

$$|\hat{S}(k)| = |Y(k)| - |\acute{N}(k)|, \quad (1)$$

где – $|\acute{N}(k)|$ – средняя амплитуда спектра шума. Величина $|\acute{N}(k)|$ вычисляется на основе предположения о локальной стационарности шума.

При этом очевидно, что при превышении мгновенного спектра неочищенной речи $|Y(k)|$ средней амплитудой спектра шума $|\acute{N}(k)|$ амплитуда спектра $|\hat{S}(k)|$ может принимать отрицательные значения. Для исключения подобных ситуаций применяется метод «выпрямления» значений $|\hat{S}(k)|$ (ограничения значений $|\hat{S}(k)|$ некоторой минимальной величиной (Noise Floor)).

Очевидно, что предельно достижимая эффективность данного метода ограничивается условием $|Y(k)| > |\acute{N}(k)|$. То есть спектральные компоненты РС, амплитуда которых на кадре обработки сопоставима с амплитудой спектра шума, уже не могут быть выделены. Это условие приблизительно соответствует наблюдаемости «следов» РС на спектрограмме.

Согласно результатам исследования возможностей нескольких методов спектрального вычитания[10] их применение для фонограмм, зашумленных белым шумом и помехой типа «речевой хор», при сегментарном отношении «сигнал\шум» (далее *SNR*) первоначальной записи 0 дБ позволяет обеспечить приращение указанного отношения на 5,5 дБ и 3 дБ соответственно.

Схожие результаты были получены в работе[11], где исследователям удалось обеспечить приращение сегментарного *SNR* на 8,5 дБ и 6,5 дБ соответственно для записей с исходным сегментарным SNR минус 10 дБ.

В работе[12] приращение указанного показателя относительно первоначального значения минус 5 дБ для фонограммы, зашумленной белым шумом, составило 14 дБ.

Самой распространенной проблемой методов спектрального вычитания является появление помех вида «музыкальный шум», т.е. возникновение в спектре очищенного РС изолированных максимумов, звучащих после преобразования сигнала во временное пространство как случайные тона. При этом многими исследователями отмечается, что «музыкальный шум» зачастую снижает восприятие РС сильнее, чем исходный стационарный шум. Поэтому значительные усилия направлены на разработку способов нивелирования влияния «музыкального шума» на разборчивость речи [1–3]. Одним из них является добавление к очищенному РС с остаточными музыкальными вставками низкоуровневого фонового шума, нивелирующего музыкальный эффект без потери речевой разборчивости (РР).

Большинство современных методов анализа звуков речи основаны на спектральной модели стационарного сигнала. Недостатком такой модели является отсутствие вероятностных характеристик для основных шумовых составляющих в произносимых согласных (консонантных фонем).

В рамках реализации алгоритмов спектрального вычитания определяются акустические признаки только вокализованных фонем (аллофонов), в консонантных фонемах анализируется только их длительность.

Эти ограничения вызваны следующими свойствами преобразования Фурье при обработке нестационарных сигналов: исходный сигнал заменяется на периодический с периодом, равным длительности анализируемого участка; преобразование Фурье не обеспечивает необходимую точность при изменении параметров процесса во времени (нестационарности), поскольку дает усредненные коэффициенты для всего исследуемого сигнала. Для выполнения анализа нестационарного процесса необходимо использовать базисные функции, имеющие способность выявлять в анализируемом сигнале как частотные, так и его временные характеристики. Другими словами, сами функции должны обладать свойствами частотно-временной локализации [4, 5]. В этой связи стоит обратить внимание на методы вейвлет-обработки.

9 Boll S., Suppression of acoustic Noise in Speech Using Spectral Subtraction, IEEE Tr. On ASSP, vol.27, N4, pp.113–120, 1979.

10 Asaduzzaman M., A Spectral Domain Speech Enhancement Method Based on Noise Compensations in Both Magnitude and Phase Spectra, http://lib.buet.ac.bd:8080/xmlui/bitstream/handle/123456789/3452/Full%20Thesis.pdf?sequence=1&isAllowed=y.

11 T. Gerkmann, C. H. Richard, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay, IEEE Transactions on Audio Speech and Language Processing, vol. 20, no. 4, pp. 1383–1393, 2012.

12 I. Cohen and B. Berdugo, Speech enhancement for non-stationary noise environments, ELSEVIER Signal Process., vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

Методы, основанные на оценке спектральных характеристик помех (методы вейвлет-обработки)

Большинство алгоритмов шумопонижения реализуется в пространстве частот с использованием кратковременного преобразования Фурье (*Short-time Fourier transform, STFT*), которое позволяет анализировать нестационарные сигналы. *STFT* реализует компромисс между временным и частотным разрешением. Однако *STFT* формирует для всех частот одинаковое разрешение по времени, что не вполне согласуется со сложной структурой фонем (аллофонов) речи. Некоторые алгоритмы шумопонижения разработаны с использованием вейвлет-преобразований, дающих более гибкое частотно-временное представление РС [6, 7].

Одним из популярных алгоритмов вейвлет шумопонижения является алгоритм вейвлет сжатия. Алгоритм вейвлет сжатия основан на сравнении вейвлет коэффициентов с заданным порогом. Оцениваемый порог задает границу между коэффициентами, соответствующими шуму и коэффициентами, соответствующими РС. Однако разделить коэффициенты, соответствующие шуму и сигналу с использованием порога не всегда возможно, особенно для консонантных фонем.

Для зашумленной речи энергия вокализованных звуков сопоставима с энергией шума. Использование одинакового порога для всех коэффициентов преобразования приводит не только к подавлению шума, но и самих вокализованных фонем речи[13]. Это приводит к плохому качеству РС после обработки.

Более удачной идеей является комбинирование вейвлет банка фильтров с фильтрацией (например, фильтрацией Винера) в пространстве вейвлет коэффициентов[14] или применения адаптивной пороговой фильтрации коэффициентов дискретного обучаемого вейвлет-преобразования [8].

Алгоритм фильтрации при помощи вейвлет-преобразования позволяет эффективно удалять высокочастотный шум, даже превышающий по величине исследуемый сигнал (следует отметить, что маскирующий эффект создают преимущественно низкочастотные сигналы), в то время как преобразование Фурье теряет информацию об особенностях низкочастотной части сигнала, что приводит к искажению временной формы полезного сигнала.

Вейвлет-преобразование отличается наиболее сложной и гибкой структурой представления сигналов в пространстве «масштаб-время». Это дает возможность более полного и тонкого вейвлет-анализа РС, по сравнению с другими известными видами анализа. Более того вейвлет-преобразование позволяет более достоверно отобразить кратковременные консонантные фонемы. При этом особенности сигналов «привязаны» к временной шкале.

Основной проблемой фильтрации при помощи вейвлет-преобразования является выбор вида материнского вейвлета для проведения анализа[15]. Очевидно, что вид вейвлета должен повторять форму исходного сигнала. В качестве материнских вейвлетов при получении частотно-временного представления РС (сонограмм) чаще всего выбирают следующие вейвлеты: Морле, Шеннона, «мексиканская шляпа», вейвлет Дебеши.

Вейвлет Морле относится к «грубым» вейвлетам и представляет сигнал с меньшей точностью, чем вейвлет Шеннона. Применение вейвлета Морле целесообразно при анализе сигналов с частотой дискретизации близкой к 8 кГц, либо сигналов, подвергнутых компрессии.

Применение вейвлета «мексиканская шляпа» целесообразно при анализе кратковременных участков сигнала, так как обеспечивается возможность «рассмотреть» каждый период сигнала в отдельности.

Вейвлет Дебеши применяют для отыскания межфонемных границ в случае, когда форма речевого тракта при переходе от аллофона к аллофону изменяется относительно медленно.

Ввиду того, что заранее невозможно предугадать форму РС и невозможно определить, на каком масштабе нужно искать интересующую нас информацию, выбор «материнского вейвлета» представляет нетривиальную задачу.

В зависимости от исходного *SNR* РС, применяемого (в рамках алгоритма) фильтра, а также вида материнского вейвлета алгоритмы вейвлет-обработки сигналов дают различный выигрыш в *SNR*. Приращение *SNR* для фонограмм с исходным *SNR* порядка 6–10 дБ составляет в среднем около 8 дБ[16].

В исследовании отмечается, что корректное сопоставление результатов работы всех методов вейвлет-обработки фактически невозможно, поскольку каждый из алгоритмов шумопонижения по-своему деформирует исходную запись и формирует в обработанном варианте свои артефакты, которые совершенно по-разному воспринимаются разными слушателями.

В рамках исследования рассчитывалось *SNR* для исходной записи голоса, зашумленной белым шумом, и прошедшей обработку по анализируемому алгоритму. Для анализируемых фонограмм методы,

13 H. Tasmaz, and E. Ercalebi, Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments, Digital Signal Process., vol.18. N.5. pp.797–812, 2008.

14 M. K. Hasan, S. Saluhuddin, and M. R. Khan, Reducing signal-bias from mad estimated noise level for dct speech enhancement, Signal Process., vol.84. N.1. pp.151–162, 2004.

15 Горшков Ю. Г. Обработка речевых сигналов на основе вейвлетов // T-comm: Телекоммуникации и транспорт. – 2015. – № 2. – С. 46–53.

16 Wieland B. Speech Signal noise reduction with wavelets, https://www.uni-ulm.de/fileadmin/website_uni_ulm/uzwr/wieland/wieland-diplomarbeit-speech-signal-noise-reduction-with-wavelets.pdf.

основанные на БПФ и стационарном вейвлет-преобразовании (*SWT – Stationary Wavelet Transform*), продемонстрировали меньшую эффективность, нежели методы быстрого вейвлет-преобразования (*FWT – Fast Wavelet Transform*).

Однако в исследовании[17] отмечается, что при малых исходных значениях *SNR* фонограмм (10 дБ и менее) все указанные алгоритмы значительно снижают уровень консонант (например, в одной из фонограмм после применения алгоритма оказался полностью вырезанным звук «т»).

В исследовании [9] также отмечается, что классические подходы к улучшению качества речи, основанные на задании порога в области вейвлет-преобразования, могут вносить определенные искажения в исходный РС. Особенно это касается глухих консонант. Поэтому зачастую указанные методы комбинируют с другими, такими как: спектральное вычитание, Винеровская фильтрация и т.д.

В исследовании [9] предложен метод шумопонижения, основанный на оценке минимального среднеквадратического отклонения значений сигнала, прошедшего процедуру шумопонижения, от исходного сигнала в пространстве вейвлетов, который демонстрирует заметно большую эффективность нежели классические методы спектрального вычитания *MMSE (Minimum Mean Square Error)* и *MMSE-SMPO (Minimum Mean Square Error Soft Masking Based on Posteriori SNR Uncertainty)*. Приращение *SNR* для фонограммы, зашумленной белым шумом, с исходным *SNR* минус 5 дБ составило 13,4 дБ[18].

Простые методы фильтрации

Как правило, методы фильтрации, основанные на оценке спектральных характеристик помех, эффективны только при попытке устранения стационарной помехи, спектральный состав которой заранее известен. Такие фильтры предназначены для компенсации в РС достаточно узкополосных квазистационарных помех. Практическими примерами таких помех могут служить промышленные помехи сети электропитания, трансформаторные шумы, сосредоточенные по спектру шумы механизмов и т.п.

Очевидно, что такой алгоритм наиболее работоспособен в условиях стационарной периодической помехи и является оптимальным для фильтрации гармонической помехи, снижение уровня которой в случае применения указанного метода шумопонижения возможно на величину до 25–35 дБ[19].

В то же время даже в случае заранее известного спектрального состава помехи при применении фильтрации происходит значительное искажение полезного сигнала, связанное с вычитанием спектра в полосе частот, занимаемой помехой. В случае помехи с узкополосным спектром, близкой к гармонической, сужение полосы режекции вызывает появление боковых лепестков (явление Гиббса). Правильный выбор вида оконной функции может снизить уровень боковых лепестков, но ценой ухудшения разрешающей способности по частоте и не связанным с этим искажением исходного сигнала.

Наиболее часто в процедурах шумопонижения используют окно Хемминга или усеченное окно Гаусса с минимальным уровнем боковых лепестков в частотной области.

Методы коррекции и сглаживания спектра РС

Методы используют свойство периодичности вокализованной речи, в которой звонкие звуки можно представить сигналами с периодом, кратным частоте основного тона голоса (ЧОТ). Это означает, что их энергетический спектр сосредоточен в определенных полосах частот, в то время как уровень энергии спектра, связанный с воздействием искажающих факторов, в общем случае, определен по всему диапазону частот. Существующие способы реализации этого метода можно разделить на два различных вида.

Первый – это фильтрация исходного искаженного сигнала гребенчатым фильтром. Качество очистки РС зависит от точности определения ЧОТ. Поскольку она постоянно меняется, при обработке разных участков речи требуется постоянная адаптивная подстройка фильтра, что не всегда просто реализуется на практике. Например, такая фильтрация совершенно неприемлема в случае воздействия на сигнал суммы гармонических и (или) речеподобных помех.

При некорректном задании АЧХ гребенчатого фильтра (ГФ) и (или) при любом способе округления ЧОТ максимумы АЧХ ГФ уже на второй и последующих гармониках достаточно сильно отстоят от максимумов спектрального распределения исходного сигнала. Это означает, что на этих гармониках полезный сигнал не выделяется, а подавляется. Более того, выделяются спектральные составляющие помех в окрестности гармонических составляющих.

В реальном сигнале ЧОТ изменяется во времени и, следовательно, простая модель цифрового ГФ становится неэффективной[20] .

Второй способ реализации основан на совмещении процедуры оценки и фильтрации. В этом

17 N. A. Whitmal, J. C. Rutledge, J. Cohen Wavelet-based noise reduction, Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on 5:3003-3006 vol.5, 1995 DOI:10.1109/ICASSP.1995.479477.

18 Y. Lu and P. C. Loizou, «Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty» IEEE Transactions on Audio Speech and Language Processing, vol. 19, no. 5, pp. 1123–1137, 2011.

19 Дворянкин С.В. Цифровая шумоочистка аудиоинформации. – М.: ИП Радиософт. – 2011. – 208 с.

20 Чесноков М., Цифровой гребенчатый фильтр с линией задержки продолжительностью в дробное число отсчетов Научно-технические ведомости СПбГПУ 5' (181) 2013 Информатика. Телекоммуникации. Управление. – С. 9–15.

случае используется вариант адаптивного фильтра, рассмотренный в работе[21]. При этом выходной сигнал снимается с выхода компенсатора, а задержка выбирается исходя из времени корреляции РС. Полученный алгоритм фильтрации РС работает в условиях широкополосного некоррелированного шума. Частотная характеристика такого фильтра представляет собой характеристику гребенчатого фильтра.

Определенно, сложная структура РС и нестационарность процесса речеобразования значительно снижают эффективность метода и приводят лишь к незначительному повышению разборчивости.

В работе[17] приведен пример цифровой реализации гребенчатого фильтра с интерполяцией по двум отсчетам. В исследовании продемонстрирована возможность улучшения *SNR* при тестировании ГФ на обработке сигнала в аддитивной смести с белым шумом на 4–10 дБ. Однако отмечается, что величина квазистационарных интервалов в вокализованной речи и время переходных процессов фильтра не допускают применения коэффициентов обратной связи больше 0,6, в связи с чем среднее улучшение *SNR* при применении ГФ ограничивается лишь 6 дБ.

Также существуют методы шумопонижения РС, основанные на периодичности вокализованных участков речи, смысл которых заключается в фильтрации верхних частот с последующим клиппированием. Повышение разборчивости осуществляется за счет выделения и усиления частотных компонент РС, несущих основную информацию о его формантной структуре.

Поскольку помеха уменьшает модуляционную глубину исходного РС, то повысить разборчивость речи можно путем её искусственного увеличения в определенном диапазоне частот. Экспериментальное исследование показало, что некоторое повышение разборчивости можно получить увеличением модуляционной глубины РС до искажения, однако такое улучшение разборчивости незначительно и на восприятие РС практически не влияет.

Методы временной обработки, основанные на модели РС

Большая часть описанных выше методов улучшения качества РС направлены на подавление шума. Эти методы нарушают спектральный баланс РС, что сопровождается его искажением.

В работах [10, 11, 12] предложен метод улучшения качества РС, использующий характеристики источника речи, в частности выходной сигнал фильтра линейного предсказания (ФЛП). Основа подхода к повышению качества РС заключается в выявлении в зашумленной речи участков с большими SNR и сохранении этих участков неизменными в отличие от участков РС с малыми значениями SNR. Отсчеты выхода ФЛП умножаются на весовую функцию, и модифицированный остаточный сигнал подается на вход полюсного фильтра, на выходе которого синтезируется улучшенный РС.

В работе[22] предложен алгоритм, отличие которого от базового алгоритма заключается в том, что весовая функция для остатков ФЛП конструируется с использованием критерия оптимизации. Очищенный РС синтезируется на выходе изменяющего во времени характеристики полюсного фильтра, на вход которого поступают остатки ФЛП.

В работе[23] предложено использовать огибающую преобразования Гильберта для реконструирования взвешенного сигнала остатков ФЛП. Огибающая преобразования Гильберта является хорошим индикатором вокализованных звуков. Поэтому применение к остаткам ФЛП преобразования Гильберта в качестве весовой функции приводит к контрастированию периодической структуры тональной речи.

В исследовании[24] проведена оценка эффективности метода линейного предсказания, предназначенного для подавления стационарных помех. Для фонограмм с различными исходными SNR (минус 10...0 дБ) в среднем приращение значения *SNR* составило около 10 дБ.

В работе[25] предложен алгоритм на основе медианного фильтра и фильтров краткосрочного (STP) и долгосрочного (LTP) линейного предсказания. Медианный фильтр в данном алгоритме предназначен для удаления нестационарных шумов в РС. Поскольку медианный фильтр сохраняет только медленно меняющиеся компоненты входного сигнала, он может исказить характеристику быстро меняющейся области речи.

Следовательно, необходим дополнительный этап предварительной обработки для сохранения характеристик речи перед применением медианного фильтра. На этапе предварительной обработки используется фильтр краткосрочного линейного предсказания (STP) и фильтр долгосрочного предсказания (LTP).

Приращение значения сегментарного SNR для исходных фонограмм с исходными *SNR* -4 дБ и -5 дБ составило 10 дБ. Однако следует отметить, что в источнике отсутствуют сведения о типе помех, которыми «зашумлялись» исходные фонограммы.

Всем методам линейного предсказания, построенным на предположении о линейности передаточной функции голосового тракта и возможности представления РС в любой произвольный момент времени

21 Дворянкин С. В. Цифровая шумоочистка аудиоинформации. – М.: ИП Радиософт. – 2011. – 208 с.

22 W. Jin, and M. S. Scordils, Speech enhancement by residual domain constrained optimization, Speech Communication, vol.48, pp.1349–1364, 2006.

23 B. Yegnanarayana et al., Speech enhancement using excitation source information, Proc. Int. conf. ICASSP, pp.1541–1544, 2002.

24 A. Kawamura, K. Fujii, Y. Itoh and Y. Fukui, A new noise reduction method using linear prediction error filter and adaptive digital filter, 2002 IEEE International Symposium on Circuits and Systems (ISCAS), Phoenix-Scottsdale, AZ, USA, 2002, pp. III-III, doi: 10.1109/ISCAS.2002.1010267.

25 Choi M.S., Kang H.G. Transient noise reduction in speech signal with a modified long-term predictor. EURASIP J. Adv. Signal Process. 2011, 141 (2011). https://doi.org/10.1186/1687-6180-2011-141

в виде линейной комбинации своих значений в предыдущие моменты, присущ один серьезный недостаток. В случае обработки сильно зашумленных РС точность вычисления коэффициентов линейного предсказания уменьшается. Это, в свою очередь, может еще сильнее ухудшить разборчивость сигнала на выходе системы линейного предсказания.

Методы адаптивного подавления помех

Методы адаптивного подавления обрабатывают параллельно искаженный сигнал и некий опорный сигнал [13]. При этом в качестве опорного сигнала используется сигнал, получаемый от датчиков, располагаемых в точках, где уровень речевого сигнала мал, то есть опорный сигнал максимально коррелирован с сигналом помехи. Следует отметить, что вычитание спектрограмм сигналов смеси и помех возможно как в режиме реального времени, так и в режиме отложенного анализа.

На первом этапе работы алгоритма создается оценка компонента, коррелированного с опорным сигналом, после чего он вычитается из смеси сигнала с шумом.

Метод адаптивного подавления реализован в двух типах систем, которые различаются способом получения опорного сигнала. В одноканальных системах (первый тип) опорный сигнал формируется из зашумленного с помощью различных преобразований последнего. В двухканальных системах используются два слабо коррелированных между собой источника смеси сигнала и шума.

Такой метод реализуют в режиме стереозаписи с двух микрофонов, которые находятся в разных точках пространства и по-разному ориентированы на источник звука.

Двухканальные системы сложны в реализации, однако, при корректных условиях размещения микрофонов и последующей правильной генерации опорного сигнала, двухканальные системы обеспечивают восстановление разборчивости даже крайне зашумленных сигналов.

Очевидно, что возможности данного метода ограничены возможностью получения опорного сигнала, максимально соответствующего помехе в искаженном сигнале, или – в случае вычитания помехи – образца помехи. В ряде случаев их удается получить позже, но тогда возникает достаточно сложная задача синхронизации сигналов основного и опорного каналов.

Методы реконструкции гармонической структуры спектральных описаний речи, искаженной шумами и помехами

Как известно, смысловая часть информации содержится в частотной огибающей РС, а основой конструкций РС являются вокализованные участки речи. Как правило, при наличии помех высшие гармоники, как наименее мощные, скрываются под шумом, а несколько мощных первых гармоник в низкочастотной области спектра с наибольшей амплитудой проявляются на фоне шумов. По этим оставшимся следам гармоник возможно нахождение значения частоты основного тона, восстановление гармонической структуры и, в итоге, звучания искаженного звукового сигнала.

В работе [14] показано, что имея в распоряжении только часть информации о гармонической структуре спектрограммы защищаемого речевого сигнала, другую её часть можно восстановить или реконструировать на основе свойств самого РС. Например, используя свойство кратности гармоник основного тона по оси частот спектрограммы. Также в указанной работе рассмотрен вариант шумоочистки, основанный на обнаружении «следов» помехи на изображении спектрограммы, синтезе помехи по найденным «следам» и её вычитании из исходной смеси.

В исследовании[26] представлены результаты параболической коррекции линий гармоник по вершинам парабол спектральных разверток. После проведенной коррекции достроенные высшие гармоники стали непрерывными и совпали с исходными зашумленными. В этой же работе отмечается, что при зашумлении РС белым шумом треки гармоник и восстановленная по ним гармоническая структура находится корректно даже при *SNR* до минус 12 дБ.

Несомненным достоинством методов реконструкции РС является то, что они работают даже в случае сильно зашумленного сигнала при условии, что треки нескольких первых гармоник различимы на фоне шумов. Однако в случае отсутствия треков первых гармоник или отсутствия возможности различения треков первых гармоник и других квазигармонических «следов» речеподобной помехи на сонограмме сигнала указанный метод не может быть успешно реализован.

Экспериментальные исследования эффективности средств шумопонижения

Для экспериментальных исследований авторами были выбраны типовые программные средства шумоочистки «Sound Cleaner», «GritTec's Noise Cancellation» и «Лазурь – М», в которых реализованы основные алгоритмы шумопонижения, рассмотренные ранее. В качестве тестовых РС были выбраны аудиозаписи таблиц слов (ГОСТ 16600-72) и фраз из (ГОСТ Р 50840-95).

Запись исходных тестовых РС проводилась в служебном помещении при уровне шума не более 35–40 дБ бригадой дикторов, не имевших дефектов

26 Дворянкин С. В., Алюшин В. М. Метод реконструкции гармонической структуры спектральных описаний искаженной шумами и помехами речи. М.: Известия института инженерной физики. – 2013. – т. 2. – № 28. – 57–62 с.

речи. Чтение слов и фраз дикторами осуществлялось ровным голосом, четко, но без подчеркивания отдельных звуков с постоянным уровнем речи. Дикторы выдерживали постоянный ритм речи на протяжении чтения всей таблицы.

Для формирования зашумленных тестовых РС использовалась ПЭВМ и специальное программное обеспечение Adobe Audition 1.5.

Для зашумления исходных тестовых РС использовались три наиболее эффективные помехи: шум со спектром, близким к нормированному спектру речи, помеха типа «речевой хор», а также «уличный» шум. Аудиофайлы формировались для *SNR* от –20 дБ до +10 дБ с шагом в 2 дБ.

Усредненные спектры тестовых сигналов в смеси с соответствующими помехами приведены на рисунках 2–4.

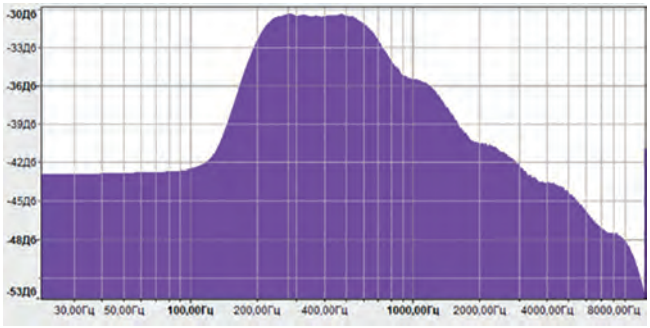


Рис. 2. Усредненная спектрограмма РС, зашумленного шумом со спектром, близким к нормированному спектру речи

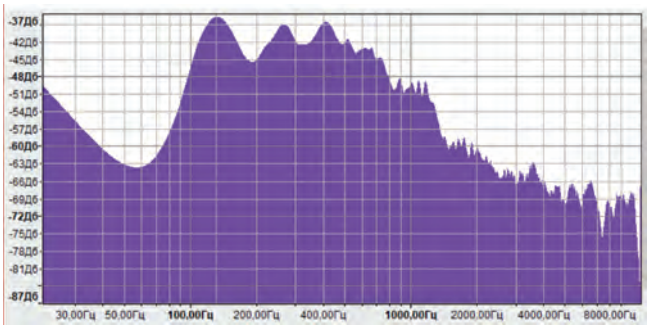


Рис. 3. Усредненная спектрограмма РС, зашумленного помехой типа «речевой хор»

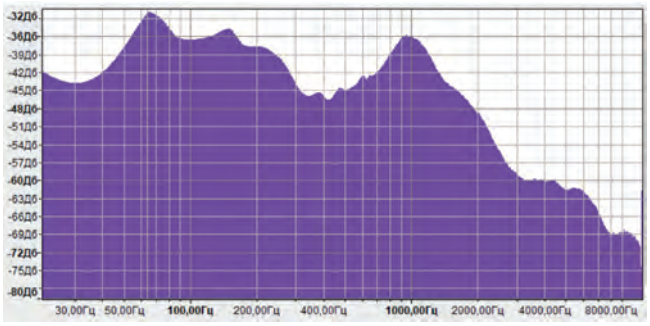


Рис. 4. Усредненная спектрограмма РС, полученного в условиях «уличного» шума

Из приведенных спектрограмм видно, что в диапазоне частот выше 1 кГц все помехи имеют спектр, близкий к «коричневому» шуму. При этом помехи типа «речевой хор» и «уличный шум» имеют локальные максимумы в диапазоне от 50 Гц до 1 кГц. Наличие указанных максимумов обусловлено, во-первых, наличием в указанном частотном диапазоне максимума нормированного спектра человеческой речи, и, во-вторых, фоновыми шумами (в большинстве своём индустриальными), которые имеют частотно зависимый характер и, как правило, уменьшаются по мере роста частоты.

Оценка разборчивости РС после шумопонижения проводилась артикуляционным методом путем оценки словесной разборчивости речи (отношения количества правильно понятых слов к их общему количеству в таблице).

Некоторые результаты исследования представлены на рис. 5.

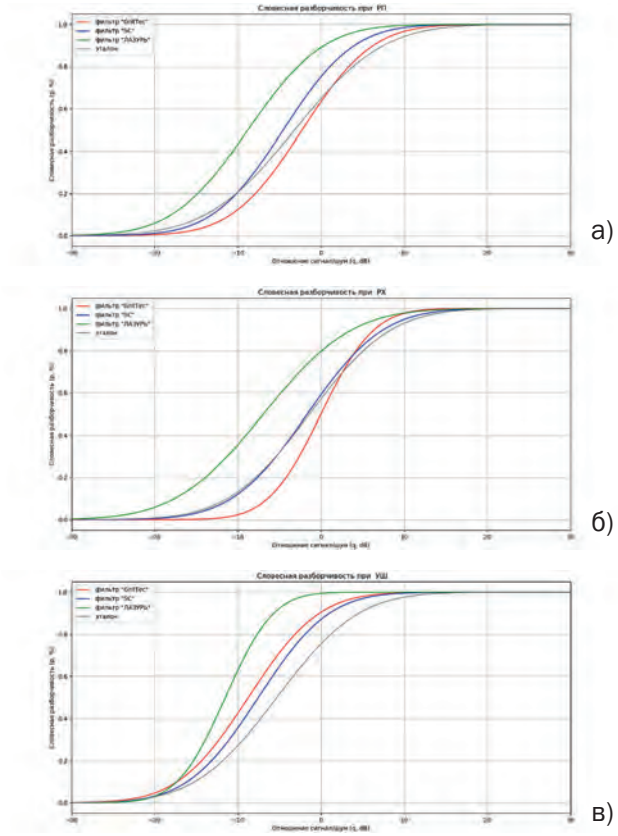


Рис. 5. Графики зависимости словесной разборчивости речи от SNR для исходной записи («эталон») и записей очищенными различными средствами шумоочистки (для средних значений) с помехами: шум со спектром, близким к нормированному спектру речи (а), «речевой хор» (б) и «уличный шум» (в) для средств шумопонижения: «GritTec's Noise Cancellation», «Sound Cleaner» и «Лазурь»

Исходя из зависимостей, представленных на рисунке 5, можно сделать вывод, что для аудиозаписей, полученных в условиях шума со спектром, близким

к нормированному спектру речи, наиболее эффективным является средство шумоочистки «Лазурь», обеспечивающее увеличение словесной разборчивости речи в области низких отношений сигнал/шум (от -16 до -6 дБ) на 22–36%. Несколько меньшую эффективность демонстрирует средство шумоочистки «Sound Cleaner». Применение средства шумоочистки «GritTec's Noise Cancellation» приводит к снижению словесной разборчивости речи.

Применение для аудиозаписей, полученных в условиях помехи типа «речевой хор», средства шумоочистки «Лазурь» обеспечивает увеличение словесной разборчивости речи в области низких отношений сигнал/шум (от -14 до -6 дБ) до 25–40%. Средство шумооочистки «Sound Cleaner» понижает словесную разборчивость при отношении сигнал/шум ниже -4 дБ, и незначительно повышает разборчивость при отношении сигнал/шум выше -4 дБ. Применение средства шумоочистки «GritTec's Noise Cancellation» приводит к снижению словесной разборчивости речи.

При этом наилучшие результаты шумоочистки аудиозаписей, полученных в условиях «уличного шума», демонстрирует средство шумоочистки «Лазурь», обеспечивающее увеличение словесной разборчивости речи в области низких отношений сигнал/шум (от -16 до -6 дБ) на 8–28%.

Проведенный эксперимент демонстрирует, что работа некоторых алгоритмов шумоочистки с сигналами может приводить к некоторому снижению разборчивости речи. Кроме того, по результатам анализа зависимостей словесной разборчивости речи от отношения «сигнал/шум», представленных на рис. 6, становится очевидным, что помеха типа «речевой хор» демонстрирует большую устойчивость по отношению к применяемым механизмам шумопонижения, нежели шум со спектром, близким к нормированной речи, и «уличный шум».

Стоит отметить, что на рисунке 5 представлены аппроксимированные функции зависимости словесной разборчивости речи от SNR. Представленные на указанном рисунке значения словесной разборчивости ввиду наличия аппроксимации несколько отличаются от фактически полученных в ходе эксперимента значений.

Таблица 1

Оценка словесной разборчивости речи и эффективности средств шумоочистки для шума со спектром, близким к нормированному спектру речи

№ п/п	Отношение сигнал/шум, дБ	Разборчивость без шумоочистки, %	Эффективность шумоочистки ΔW, %		
			«GritTec»	«Sound Cleaner»	«Лазурь»
1	- 20	0	0	0	0
2	- 18	0	0	1	2
3	- 16	2	-2	-1	5

Таблица 2

Оценка словесной разборчивости речи и эффективности средств шумоочистки для помехи типа «речевой хор»

№ п/п	Отношение сигнал/шум, дБ	Разборчивость без шумоочистки, %	Эффективность шумоочистки ΔW, %		
			«GritTec»	«Sound Cleaner»	«Лазурь»
1	- 20	0	0	0	0
2	- 18	0	0	0	0
3	- 16	0	0	0	6

Таблица 3

Оценка словесной разборчивости речи и эффективности средств шумоочистки для «уличного шума»

№ п/п	Отношение сигнал/шум, дБ	Разборчивость без шумоочистки, %	Эффективность шумоочистки ΔW, %		
			«GritTec»	«Sound Cleaner»	«Лазурь»
1	- 20	0	3	0	0
2	- 18	1	5	0	0
3	- 16	3	8	7	7

Обратим внимание на фактически полученные в ходе эксперимента значения словесной разборчивости РС без применения и с применением средств шумоочистки к аудиозаписям с отношением сигнал/шум менее -14 дБ (табл. 1–3).

Согласно результатам, представленным в табл. 1–3, применение различных методов шумоочистки к аудиозаписям с изначально малыми значениями отношения сигнал/шум не приводит к существенному увеличению словесной разборчивости. Приращение разборчивости на 3–5% крайне незначительно. Данные значения сопоставимы с ошибкой аппроксимации и попадают в доверительный интервал (0,95%) при построении зависимости словесной разборчивости от *SNR*.

Выводы

Необходимо отметить, что результаты работы различных методов шумопонижения и реконструкции РС, приведенные в проанализированных в ходе текущего исследования источниках, получены путем математического моделирования и вычислительных экспериментов. По результатам физического эксперимента методы шумопонижения и реконструкции РС по очевидным причинам продемонстрировали бы гораздо менее впечатляющие результаты.

Более того, стоит отметить, что важным аспектом корректной оценки полученных при оценке методов шумопонижения результатов является неоднозначность измерений среднеквадратической мощности (*RMS*) сформированных акустических сигналов, что не позволяет с должной точностью определить значения *SNR* смеси сигнала с шумом.

Для корректного расчета значений *SNR*, применительно к потоку слитной речи, необходимо вводить соответствующие поправки или лимитировать длительность пауз тестовых сигналов согласно характеристикам и параметрам слитной речи.

Ввиду сказанного очевидной становится практическая невозможность достаточно точного определения степени улучшения восприятия прошедшей процедуры шумопонижения фонограммы по приращению значений *SNR*.

Улучшение качества фонограммы, по отношению к которой применялись алгоритмы шумопонижения, можно оценить только субъективно (на слух), однако результаты оценки улучшения разборчивости в большинстве указанных исследований не приведены.

Более того, в подавляющем большинстве исследований отсутствуют сведения о диапазоне частот, в котором проводился анализ результатов работы методов шумопонижения.

Очевидно также, что прямое сравнение всех указанных методов по сведениям, приведенным в источниках, малоинформативно. Каждый из описанных методов оставляет в обработанных фонограммах свои артефакты, которые влияют на разборчивость речи ввиду субъективности восприятия, слушающего по-разному. Поэтому оценка эффективности того или иного алгоритма через объективный показатель *SNR* не дает полного представления о степени улучшения восприятия РС слушающим.

Кроме того, в рассмотренных работах не указано, каким образом проводилась оценка *SNR* для сигналов, прошедших процедуру шумопонижения.

Стоит также отметить, что нижняя граница *SNR*, значимых для задач оценки защищенности акустической речевой информации, находится в диапазоне около -25...-20 дБ.

Согласно работам Ю. С. Быкова[27] нижняя граница разборчивости РС в каналах связи для коррелированных тестов (команд) достигается при *SNR* -14 дБ. Судя по результатам, приведенным в проанализированных источниках, а также результатам экспериментальных исследований, методы шумопонижения и реконструкции РС также не могут преодолеть указанное Ю. С. Быковым пороговое значение, и при значении SNR исходной фонограммы менее -14 дБ не являются эффективными. Аналогичные выводы о невозможности применения механизмов шумопонижения к записям с низким *SNR* сделаны исследователями в [15].

В настоящее время алгоритмы быстрого вейвлет-преобразования демонстрируют большую эффективность нежели алгоритмы шумопонижения на основе быстрого преобразования Фурье или стационарного вейвлет-преобразования. При этом очевидно, что все указанные методы наилучшим образом работают с фонограммами, в которых наблюдается квазистационарный шум. При зашумлении фонограмм, например, речеподобной помехой указанные методы будут малоэффективны.

Методы фильтрации, основанные на оценке спектральных характеристик помех, могут быть эффективны только при попытке устранения стационарной помехи, спектральный состав которой заранее известен.

Методы коррекции и сглаживания спектра согласно результатам исследований демонстрируют весьма скромную эффективность. При этом существенным недостатком указанных методов шумопонижения является следующее: все алгоритмы коррекции спектра основаны на предположении, что уровень энергии спектра шума распределен по всему диапазону частот. Т.е. при применении для защиты информации речеподобной помехи указанные методы могут оказаться неэффективными.

27 Быков Ю. С. Теория разборчивости и повышения эффективности радиотелефонной связи / Ю.С. Быков. – М.: Госэнергоиздат, 1959. – 352 с.

Наиболее перспективными наравне с вейвлет-алгоритмами являются методы улучшения качества РС, использующие характеристики источника речи, в частности выходной сигнал ФЛП. Даже при достаточно низких *SNR* (порядка -10 дБ) данные методы демонстрируют относительно высокую эффективность.

При помощи методов реконструкции РС по трекам первых гармоник (следов фонообъектов) можно добиться значительного улучшения разборчивости РС (практически 100%) даже при очень маленьких значениях *SNR*. Однако в случае отсутствия треков первых гармоник на сонограмме сигнала указанный метод практически не может быть реализован.

Очевидно, что у всех рассмотренных методов шумопонижения и реконструкции РС есть частные, специфичные для каждого метода границы его применимости. Но у них также есть общее ограничение: все рассмотренные методы будут показывать весьма малую эффективность в случае применения для зашумления фонограммы корректно сформированной речеподобной помехи.

Постоянно меняющиеся частоты основного тона голосов дикторов, создающих помеху, делают практически невозможной задачу спектрального вычитания или фильтрации. Методы вейвлет-анализа и линейного предсказания в случае речеподобной помехи не способны разделить коэффициенты, соответствующие речеподобной помехе и РС. В случае применения методов реконструкции РС на сонограмме будет очень трудно выделить треки первых гармоник диктора на фоне треков гармоник голосов, формирующих речеподобную помеху.

Недостатки работы указанных алгоритмов с речеподобной помехой были продемонстрированы авторами в ходе эксперимента. Показано, что для двух программных сред «Sound Cleaner», «GritTec's Noise Cancellation» попытки применения алгоритмов шумопонижения записей с помехой типа «речевой хор» приводят к снижению словесной разборчивости речи.

Более того, в ходе практического эксперимента получены результаты, показывающие, что улучшение комфортности восприятия РС при применении процедур шумопонижения и реконструкции РС носит ограниченный характер. При хорошем качестве исходного РС после процедур шумопонижения из-за сопутствующих им искажений разборчивость речи может снижаться.

Следует также обратить внимание на спектральную структуру сигналов, прошедших процедуры шумопонижения: в них зачастую преобладают (наилучшим образом восстанавливаются) гармонические составляющие РС, т.е. вокализованные фонемы (аллофоны). Данное явление согласуется с тем фактом, что 60-70% разборчивости речи в потоке слитной речи дают именно вокализованные аллофоны.

Ввиду изложенного результаты проведенной работы по оценке возможностей методов шумопонижения и реконструкции РС, зашумляемых помехой типа «речевой хор», могут стать основой дальнейших исследований и разработок в области защиты речевой информации.

Литература

1. *Thimmaraja Y., Nagaraja B., Jayanna H., A spatial procedure to spectral subtraction for speech enhancement, Multimedia Tools and Applications volume 81, pages 23633–23647, 2022, https://doi.org/10.1007/s11042-022-12152-3.*
2. *Y. Yang, P. Liu, H. Zhou and Y. Tian, A Speech Enhancement Algorithm combining Spectral Subtraction and Wavelet Transform, 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2021, pp. 268–273, doi: 10.1109/AUTEEE52864.2021.9668622.*
3. *G. Ioannides, V. Rallis, Real-Time Speech Enhancement Using Spectral Subtraction with Minimum Statistics and Spectral Floor, 2023, https://doi.org/10.48550/arXiv.2302.10313.*
4. *A. Li, C. Zheng, R. Peng, and X. Li, On the importance of power compression and phase estimation in monaural speech dereverberation, JASA Express Lett., vol. 1, no. 1, pp. 014802, 2021.*
5. *T. Peer and T. Gerkmann, Phase-aware deep speech enhancement: It's all about the frame length, JASA Express Lett., vol. 2, no. 10, pp. 104802, 2022.*
6. *Бабурин А. В., Глущенко Л. А., Корзун А. М., Вейвлет-технологии для шумоочистки речевых сигналов в оптико-электронных каналах передачи информации, Информация и безопасность, 2023, Т. 26, вып. 1, с. 45–52, DOI 10.36622/VSTU.2023.26.1.006.*
7. *P. Kuwalek, W. Jesko, Speech Enhancement Based on Enhanced Empirical Wavelet Transform and Teager Energy Operator, Electronics 2023, 12(14), 3167; https://doi.org/10.3390/electronics12143167.*
8. *Лепендин А. А., Ильяшенко И. Д., Насретдинов Р. С., Применение обучаемого дискретного вейвлет-преобразования с адаптивными порогами для шумоочистки речевых сигналов, Высокопроизводительные вычислительные системы и технологии, том 4 (1), 2020, с. 51–57.*
9. *M. Talbi and M. S. Bouhlel, A New Speech Enhancement Technique Based on Stationary Bionic Wavelet Transform and MMSE Estimate of Spectral Amplitude Hindawi, Security and Communication Networks, vol. 2021, Article ID 9968275, 11 pages, 2021, https://doi.org/10.1155/2021/9968275.*
10. *X. Feng, N. Li, Z. He, Y. Zhang and W. Zhang, DNN-Based Linear Prediction Residual Enhancement for Speech Dereverberation, 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 541–545.*

11. *Yang Liu, Na Tang, Xiaoli Chu, Yang Yang, Jun Wang, LPCSE: Neural Speech Enhancement through Linear Predictive Coding, Audio and Speech Processing, 2022, https://doi.org/10.48550/arXiv.2206.06908.*
12. *С. М. Горошко, С. Н. Петров. Метод шумоочистки речевых сигналов на основе мел-частотных кепстральных коэффициентов с использованием фильтрации Калмана / С. М. Горошко, С. Н. Петров // Известия Гомельского государственного университета имени Ф. Скорины. – 2019. – № 6 (117). – С. 103–107.*
13. *K. Tan, Z.-Q. Wang, and D. Wang, Neural spectrospatial filtering, IEEE/ACM Trans. Audio. Speech, Lang. Process., vol. 30, pp. 605–621, 2022.*
14. *Дворянкин С. В., Дворянкин Н. С., Устинов Р. А. Речеподобная помеха, стойкая к шумоочистке, как результат скремблирования защищаемой речи. Вопросы кибербезопасности, 2022, № 5(51). DOI: 10.21681/2311-3456-2022-5-14-27.*
15. *Иванов А. В., Волков Н. А. Применение методов шумоочистки для обработки речевой акустической информации, Сборник избранных статей научной сессии ТУСУР, номер 1–3, 2021, с. 34–37.*

ОЦЕНКА РИСКОВ КИБЕРБЕЗОПАСНОСТИ ЭНЕРГЕТИЧЕСКОГО СООБЩЕСТВА МИКРОСЕТЕЙ[1]

Гурина Л. А.[2],

DOI: 10.21681/2311-3456-2024-1-101-107

Цель исследования: *разработка методического подхода оценки риска кибербезопасности микросетей со взаимосвязанными информационными системами.*

Методы исследования: *марковские процессы, вероятностные методы, методы теории нечетких множеств.*

Результат исследования: *рассмотрена иерархическая структура управления энергетическим сообществом, выявлены возможные кибератаки на систему управления сообществом микросетей, приведена классификация кибератак, последствием которых является нарушение качество информации. Предложена модель состояний информационной системы микросети, на основе которой получена структурная модели развития состояний энергетического сообщества при различных способах управления. Разработан подход оценки риска кибербезопасности информационно-коммуникационной инфраструктуры сообщества микросетей.*

Научная новизна *состоит в том, что для оценки риска кибербезопасности информационно-коммуникационной инфраструктуры сообщества микросетей при различных способах управления им в работе предложен подход, позволяющий учитывать возможные состояния информационных систем при кибератаках.*

Ключевые слова: *киберфизическая энергетическая система, микросеть, информационная система, качество информации, риск кибербезопасности, кибератаки.*

ASSESSMENT OF CYBER SECURITY RISK OF MICROGRIDS ENERGY COMMUNITY[3]

Gurina L. A.[4]

The research aims *to develop a methodological approach to assessing the cybersecurity risk of microgrids with interconnected information systems.*

The research relies *on the Markov processes, probabilistic methods, methods of fuzzy set theory.*

Research result: *The hierarchical management structure of energy communities is considered, possible cyber-attacks on the microgrid community management system are identified, and a classification of cyber-attacks is given, the consequence of which is a violation of the quality of information. A model of states of the microgrid information system is proposed, on the basis of which a structural model of the development of states of the energy community under various control methods is obtained. An approach has been developed for assessing the cybersecurity risk of the information and communication infrastructure of a microgrid community.*

The scientific novelty lies in the *fact that in order to assess the cybersecurity risk of the information and communication infrastructure of the microgrid community under various methods of managing it, the work proposes an approach that allows taking into account the possible states of information systems during cyber-attacks.*

Keywords: *cyber-physical energy system, microgrid, information system, information quality, cybersecurity risk, cyber-attacks.*

1 Работа выполнена в рамках научного проекта «Теоретические основы, модели и методы управления развитием и функционированием интеллектуальных электроэнергетических систем», № FWEU-2021-0001.

2 Гурина Людмила Александровна, кандидат технических наук, доцент, старший научный сотрудник Лаборатории управления функционированием электроэнергетических систем Института систем энергетики им. Л. А. Мелентьева СО РАН, Иркутск, Россия. E-mail: gurina@isem.irk.ru

3 The research was conducted within the framework of the scientific project «Theoretical foundations, models and methods to control the expansion and operation of intelligent electric power systems (Smart Grids)», No. FWEU-2021-0001.

4 Liudmila A. Gurina, Ph.D. in engineering, Associate Professor, Senior Research Fellow, Laboratory for Control of Electric Power Systems at Melentiev Energy Systems Institute, SB RAS, Irkutsk, Russia. E mail: gurina@isem.irk.ru

Введение

Электроэнергетические системы (ЭЭС) претерпевают радикальные изменения своих свойств не только за счет трансформации своей внутренней структуры, но и за счет использования инновационных технологий производства, передачи, хранения, распределения и потребления электроэнергии [1]. Возможности использования возобновляемых источников энергии привели все большему распространению микросетей. Микросети включают в себя такие источники распределенной генерации, как ветряные турбины, дизель-генераторы, топливные элементы, фотоэлектрические системы, системы хранения энергии и т.д. Масштабное применение силовой электроники, инверторов и других цифровых устройств при эксплуатации микросетей привели к росту их уязвимостей к киберугрозам. Несмотря на многочисленные преимущества с технической, экономической и экологической точек зрения, объединение микросетей в энергетические сообщества (ЭСО) [2] способствует появлению дополнительных кибеуязвимостей из-за расширения информационно-коммуникационной инфраструктуры и путей передачи данных в зависимости от способа управления ими. Становится важным оценка риска кибербезопасности ЭСО микросетей.

Применение информационных и коммуникационных технологий играет важную роль при эксплуатации и управлении ЭСО микросетями. Интеграция информационных систем и технологической части микросетей трансформирует их в киберфизические энергетические системы [3] с развитыми программными сетями управления и связи, что ведет к взаимозависимостям между информационной и физической инфраструктурами микросетей [4]. Неисправности и сбои в одной из подсистем могут передаваться между ними, усугубляя последствия как для микросетей, так и для всего сообщества в целом. Так, киберинцидент в информационной системе одной микросети может повлиять на работу не только подвергшейся кибератаке микросети, но и на надежное функционирование остальных микросетей в составе сообщества. Таким образом, работа ЭСО зависит от киберустойчивости взаимозависимых информационных систем микросетей. Нарушение качества информационных потоков в результате кибератак [5], например, задержка или искажение данных, может повлиять на надежную и бесперебойную работу технологической части ЭСО и, тем самым, поставить под угрозу устойчивое и безопасное функционирование интеллектуальных сетей в целом.

Современные системы SCADA, эксплуатируемые при управлении микросетями, перешли к использованию стандартных коммуникационных технологий, чтобы обеспечить доступ к удаленным устройствам и упростить интерфейс между устройствами от разных поставщиков. Следовательно, количество возможных точек атаки, которые могут использовать злоумышленники, резко возросло. Другой распространенной практикой является использование стандартных аппаратных и программных платформ для снижения затрат и повышения гибкости [6]. Новые уязвимости увеличивают риск киберугроз для большого количества SCADA-систем.

Целью работы является разработка методическая подхода оценки риска кибербезопасности сообщества микросетей с учетом взаимосвязей информационных систем.

Для повышения кибербезопасности и киберустойчивости ЭСО необходимо учитывать взаимозависимости не только между информационно-коммуникационной и технологической инфраструктурой микросети, но и сложную взаимосвязь внутри информационно-коммуникационной инфраструктуры сообщества микросетей.

Анализ кибербезопасности сообществ микросетей

А. Структура и задачи управления ЭСО микросетей.

При управлении ЭСО обычно используется иерархическая структура (рис. 1). Основная концепция различных методов управления подразделяется на три уровня: нижний, средний и верхний. Эти методы используются для обеспечения координации между микросетями, которая зависит от многих факторов, включая скоординированный контроль с сетями связи и без них. Уровни управления применимы для работы ЭСО как в сетевом, так и в изолированном режиме.

На нижнем уровне осуществляется управление локальной мощностью, напряжением и током, следуя значениям параметров, заданным контроллерами верхнего уровня. Основными переменными являются выходное напряжение, частота и отслеживаемые значения, получаемые от внутреннего контура управления. Основными целями управления являются первичное регулирование частоты, первичное регулирование напряжения, автоматической частотной разгрузки [7].

На среднем уровне управлении решаются задачи системного уровня, такие как регулирование качества электроэнергии, синхронизация микросетей в составе сообщества, координация распределенной генерации и т.д. [8]. Основными задачами среднего уровня является управление спросом, выбор состава включенного генерирующего оборудования, вторичное регулирование частоты, вторичное регулирование напряжения, прогнозирование

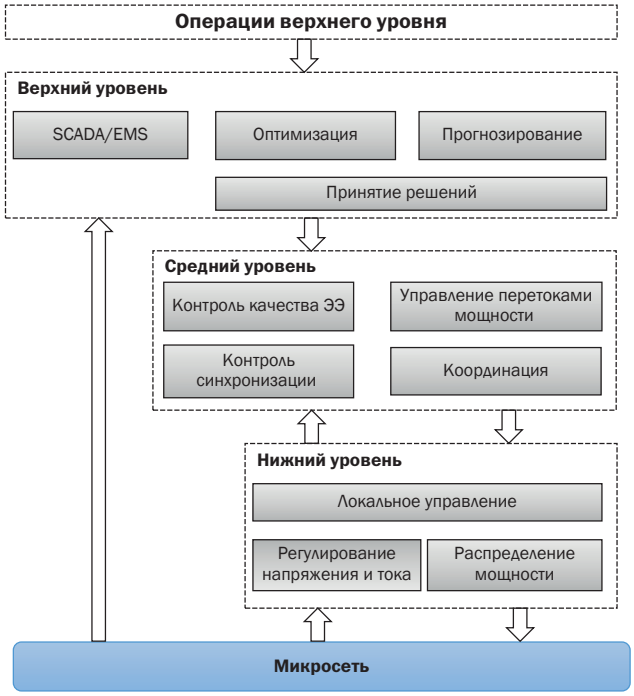


Рис. 1. Иерархическая схема управления сообществом микросетей

графиков нагрузки потребителей, прогнозирование выработки объектами распределенной генерации [9].

Решение задач оптимизации, управления и общего регулирования системы решается на верхнем уровне [10]. С технической и экономической точек зрения оптимальная работа всех генераторов достигается за счет использования методов верхнего уровня управления. Все микросети в составе ЭСО учитывают как технические, так и экономические аспекты системы управления распределением (DMS).

Нижний уровень управления в качестве базового уровня, объединяет контуры управления, направленные на регулирование напряжения, тока и мощности, а также определяет динамические характеристики локальных устройств. Средний и верхний уровни управления обеспечивают такие расширенные функциональные возможности, как поддержание качества напряжения, улучшение распределения тока и оптимизация работы.

Кибератаки могут повлиять на контроль напряжения микросетей, системы управления энергопотреблением и управление потребляемой мощностью [11].

Напряжение интеллектуальной микросети обычно контролируется распределенными генераторами с интерфейсом силовой электроники. В таких системах измеряется уровень напряжения и/или реактивная мощность системы, а система управления вырабатывает реактивную эталонную мощность для выработки электроэнергии. Атаки FDI, которые изменяют измеренное датчиком напряжение и/или данные реактивной мощности, параметры управления, могут повлиять на регулирование напряжения микросети [12]. Более того, злоумышленники могут получить доступ к многоуровневой системе управления микросетью (рис. 1) и изменить управляющие сигналы между уровнями (например, внести ошибки в опорные измерения мощности распределенной генерации).

Кибератаки, нацеленные на частоту микросетей, называются атаками на переходные процессы в микросетях. Злоумышленники могут вносить ошибки в управляющие сигналы между уровнями управления, изменять параметры управления и измерения датчиков или изменять выходные параметры источников питания, чтобы повлиять на изменение частоты микросети [13]. Следует отметить, что регулирование частоты микросети чувствительно к измерениям активной мощности и частоты, а также опорным сигналам. В микросетях частота обычно регулируется вращающимися машинами. Любые атаки, направленные на измерения скорости или угла ротора, могут повлиять на частоту микросетей. В последнее время для повышения устойчивости микросетей используются системы накопления энергии.

Б. Классификация кибератак на информационно-коммуникационную инфраструктуру энергетического сообщества

Первичный и вторичный уровни управления, несущие важную информацию, подвержены кибератакам. Кибератаки в микросетях вызывают не только проблемы с целостностью, доступностью и конфиденциальностью данных, но и могут привести таким к неблагоприятным последствиям как нарушение управления, отказы функционирования всего энергетического сообщества.

Передача данных от измерительных устройств и обмен данными между информационными системами микросетей необходимы для достижения эффективного управления сообществом микросетей. Непрерывный мониторинг и анализ данных играет важную роль в обеспечении качества информации, используемой при управлении микросетями.

Кибератаки на информационно-коммуникационную инфраструктуру можно разделить на атаки на целостность, доступность и конфиденциальность данных. Атака на целостность – это кибератака, последствием которой является недостоверность информации. Наиболее распространенной из кибератак на целостность является атака внедрения ложных данных (FDI-атака). Атака на доступность – это кибератака, которая препятствует своевременному получению необходимых данных или сигналов. К этому типу кибератак относятся атаки отказа

Таблица 1

Классификация кибератак, нарушающих качество информационных потоков при управлении сообществом микросетей

Целостность	Доступность	Конфиденциальность
FDI-атака	Jamming-атака	Социальная инженерия
Hijacking-атака	Wormhole-атака	Подслушивание
Подделка данных	DoS-атака	Анализ трафика Несанкцио-нированный доступ
Атака повторного воспроизведения	DDos-атака	Кража паролей
Wormhole-атака	Переполнение буфера Puppet-атака	Атака «Человек посередине»,
Spoofing-атака	Time Synchronization	Атака перехвата
Атака модификации	Masquerade -атака	Атака повторного воспроизведения
Атака «Человек посередине»	Атака «Человек посередине»	Masquerade -атака
Masquerade-атака	Spoofing-атака	

в обслуживании (DoS-атака). Атака на конфиденциальность относится к кибератакам, при которых неавторизованные лица незаконно получают информацию. Как правило, атаки на конфиденциальность не затрагивают систему напрямую, но часто сочетаются с другими атаками. В таблице 1 приведены возможные кибератаки на информационно-коммуникационную инфраструктуру энергетического сообщества, направленные на целостность, доступность и конфиденциальность информации [14–28].

Оценка рисков кибербезопасности информационно-коммуникационной инфраструктуры сообщества микросетей

На основе описанной иерархии управления способ реализации уровней управления ЭСО микросетей может быть централизованным, децентрализованным или распределенным (рис. 2) [29].

При кибератаках на информационно-коммуникационную инфраструктуру энергетического сообщества возможные состояния информационной системы i-й микросети можно охарактеризовать на основе двухуровневой модели:

$$S_i = \begin{cases} 1, \textit{информационная система микросети подвержена кибератаке} \\ 0, \textit{в противном случае}. \end{cases} \quad (1)$$

Кибератаки на ЭСО могут быть направлены как информационную систему одной микросети, так и на информационные системы нескольких микросетей. Также, в зависимости от способа управления сообществом управления с учетом взаимовлияния информационных систем микросетей, последствия кибератаки на одну из микросетей может быть распространено на информационные системы других микросетей. В [30] при моделировании кибератак на информационные системы микросетей при распределенном вторичном управлении сообществом микросетей проведен

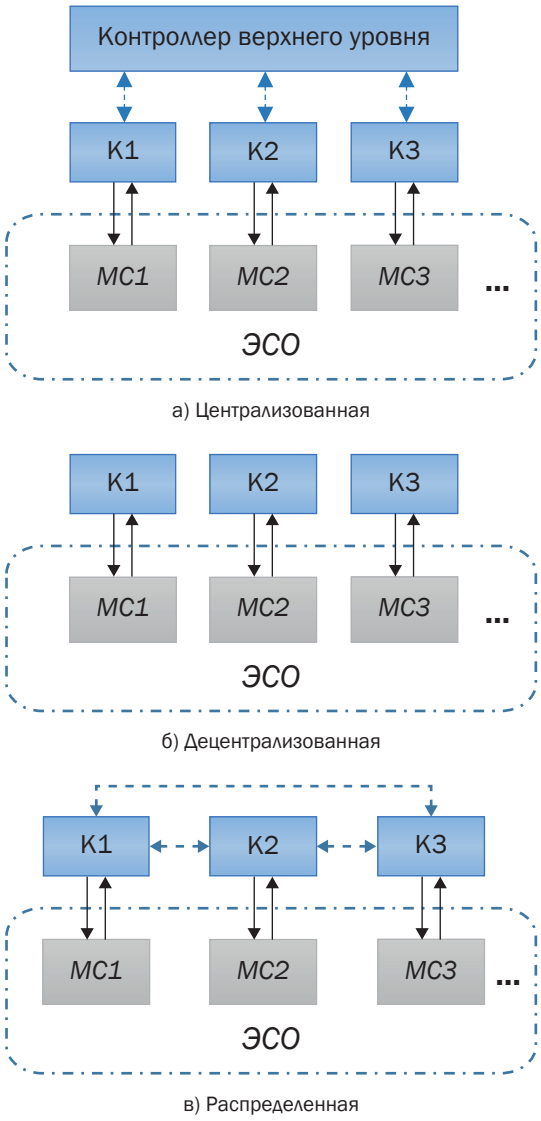


Рис. 2. Способы реализации уровней управления (- - - - - - – потоки данных при взаимодействии микросетей в составе сообщества и управлении им)

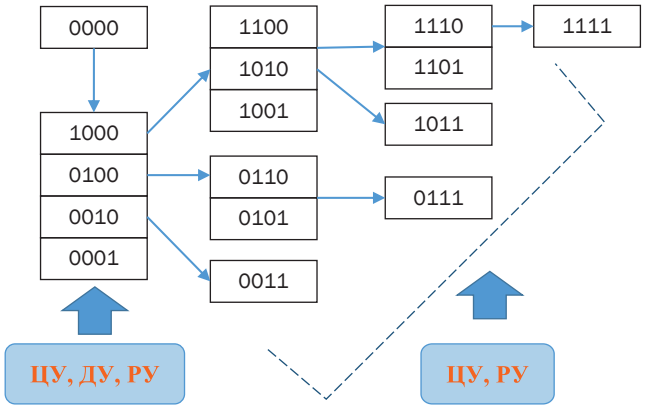


Рис. 3. Возможные состояния информационно-коммуникационной инфраструктуры ЭСО при кибератаках (ЦУ – централизованное управление, ДУ – децентрализованное управление, РУ – распределенное управление)

анализ последствий кибератак на остальные информационные системы микросетей. Наиболее опасными по последствиям для информационно-коммуникационной инфраструктуры энергетического сообщества являются FDI-атака и Hijacking-атака.

На примере четырех микросетей в составе сообщества на основе марковских процессов смоделированы возможные состояния информационно-коммуникационной инфраструктуры при различных способах управления – централизованном, децентрализованном и распределенном (рис. 3.).

Оценка риска кибербезопасности информационно-коммуникационной инфраструктуры ЭСО может быть проведена на основе следующей нечеткой модели

$$\tilde{R}_S = \prod_{i=1}^{N} \tilde{R}i \quad , \qquad (2)$$

где $\tilde{R}_i$ – уровень риска кибербезопасности i-й микросети, N – количество микросетей в составе сообщества.

Оценка уровня риска кибербезопасности i-й микросети определяется при использовании разработанной в [31] иерархической нечеткой системы (рис. 4).

Согласно модели (2) разработана иерархическая нечеткая система определения риска кибербезопасности сообщества микросетей. Для описанного примера ЭСО, включающего четыре микросети, на рис. 5. представлена нечеткая система оценки риска кибербезопасности. Семантическое описание уровней риска кибербезопасности информационно-коммуникационной инфраструктуры сообщества микросетей дано в табл. 2.

Заключение

Выявлены возможные кибератак на информационно-коммуникационную инфраструктуру энергетического сообщества микросетей. Приведена классификация кибератак, последствием которых является нарушение качество информационных потоков. Предложена модель состояний информационных систем микросетей в составе сообщества, на основе которой получена структурная модель развития состояний взаимосвязанных информационных систем микросетей в составе ЭСО. Разработан подход для оценки риска кибербезопасности информационно-коммуникационной инфраструктуры сообщества микросетей.

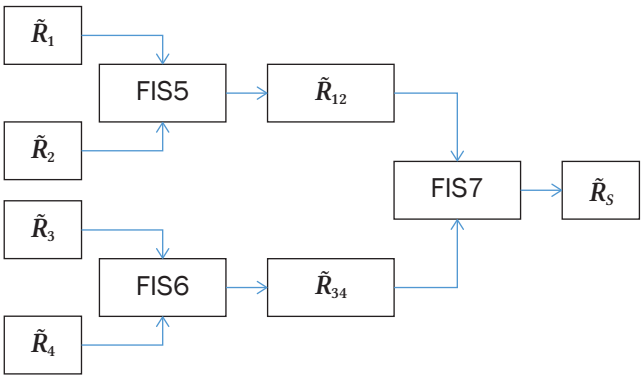


Рис. 5. Иерархическая нечеткая система оценки риска кибербезопасности сообщества микросетей

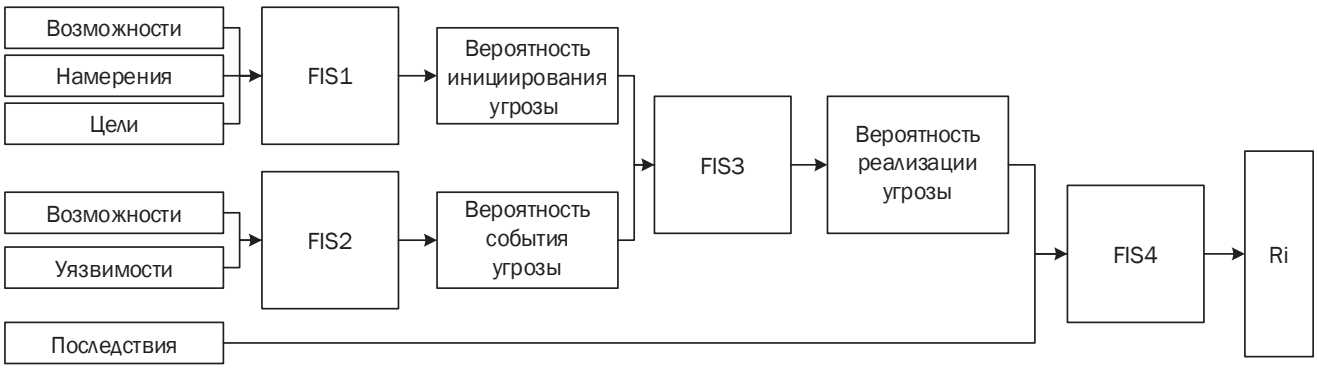


Рис. 4. Иерархическая нечеткая система оценки рисков кибербезопасности микросети

Таблица 2

Уровни риска кибербезопасности ИС микросетей

Уровень/ диапазон изменения	Описание
Низкий L,[0,0.24]	Реализованная кибератака на микросеть не приводит к отказам и сбоям компонентов ИС как самой микросети, так и компонентов ИС остальных микросетей в составе сообщества. Срабатывают все меры по обеспечению киберустойчивости. Функциональность системы управления высокая.
Средний M,[0.25,0.74]	В результате кибератаки на микросеть возможны незначительные сбои и ошибки в управлении сообществом микросетей, которые устранимы и не оказывают критического влияния на функциональность информационно-коммуникационной инфраструктуры. Реализация функций управления осуществляется в требуемом объеме и не приводит к нарушениям функциональности сообщества микросетей.
Высокий H,[0.75,1]	Опасность возникновения отказов и сбоев в энергетическом сообществе высокая в результате кибератак на ИС микросети. Сочетание отказов компонентов и/или ошибок в информационно-коммуникационной инфраструктуре может привести к значительным нарушениям функционирования сообщества микросетей.

Литература

1. *Voropai N. Electric Power System Transformations: A Review of Main Prospects and Challenges. Energies. 2020, 13, 5639. DOI: 10.3390/en13215639*
2. *Gjorgievski V. Z., Cundeva S., Georghiou G. E. Social arrangements, technical designs and impacts of energy communities: A review // Renewable Energy. 2021, vol. 169, pp. 1138–1156. DOI: 10.1016/j.renene.2021.01.078.*
3. *Zografopoulos Ioannis, Ospina Juan, Liu XiaoRui, Konstantinou, Charalambos. Cyber-Physical Energy Systems Security: Threat Modeling, Risk Assessment, Resources, Metrics, and Case Studies. 2021*
4. *H. Pan, H. Lian, C. Na and X. Li. Modeling and Vulnerability Analysis of Cyber-Physical Power Systems Based on Community Theory // in IEEE Systems Journal. Sept. 2020, vol. 14, no. 3, pp. 3938–3948. DOI: 10.1109/JSYST.2020.2969023.*
5. *Колосок И. Н., Гурина Л. А. Идентификация кибератак на системы SCADA и СМПР в ЭЭС при обработке измерений методами оценивания состояния // Электричество. 2021, №6, с. 25–35. DOI:10.24160/0013-5380-2021-6-25-32*
6. *D. Pliatsios, P. Sarigiannidis, T. Lagkas and A. G. Sarigiannidis. A Survey on SCADA Systems: Secure Protocols, Incidents, Threats and Tactics // in IEEE Communications Surveys & Tutorials. 2020, vol. 22, no. 3, pp. 1942–1976. DOI: 10.1109/COMST.2020.2987688.*
7. *Unamuno E., Barrena JA. Equivalence of primary control strategies for AC and DC microgrids // Energies. 2017, 10(1), pp. 1–13.*
8. *Jin C, Wang J, Wang P. Coordinated secondary control for autonomous hybrid three-port AC/DC/DS microgrid // CSEE J Power Energy Syst. 2018, 4(1), pp. 1–10.*
9. *Илюшин П. В. Перспективы развития и принципы построения систем автоматическоо управления режимами микроэнергосистем // Материалы юбилейной X Международной научно-технической конференции «Электроэнергетика глазами молодежи-2019». Том 1, 2019, с. 59–64.*
10. *Zakariazadeh A, Jadid S, Siano P. Smart microgrid energy and reserve scheduling with demand response using stochastic optimization // Int J Electr Power Energy Syst. 2014, 63, pp. 523–533.*
11. *Sahoo S., Mishra S., Peng, J. C., Dragicevic T. A Stealth Cyber Attack Detection Strategy for DC Microgrids // IEEE Trans. Power Electron. 2019, 34, pp. 8162–8174.*
12. *Hao J., Kang, E., Sun J., Wang Z., Meng, Z., Li X., Ming Z. An Adaptive Markov Strategy for Defending Smart Grid False Data Injection from Malicious Attackers // IEEE Trans. Smart Grid. 2018, 9, pp. 2398–2408.*
13. *Chen C., Zhang, K., Yuan K., Zhu L., Qian M. Novel Detection Scheme Design Considering Cyber Attacks on Load Frequency Control // IEEE Trans. Ind. Inform. 2018, 14, pp. 1932–1941.*
14. *M. Z. Gunduz, R. Das. Analysis of cyber-attacks on smart grid applications // in: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). 2018, pp. 1–5. DOI:10.1109/IDAP.2018.8620728*
15. *H. Zhang, B. Liu and H. Wu. Smart Grid Cyber-Physical Attack and Defense: A Review // in IEEE Access. 2021, vol. 9, pp. 29641–29659. doi: 10.1109/ACCESS.2021.3058628*
16. *V. S. Rajkumar, A. Ştefanov, A. Presekal, P. Palensky and J. L. R. Torres. Cyber Attacks on Power Grids: Causes and Propagation of Cascading Failures // in IEEE Access. 2023, vol. 11, pp. 103154–103176. DOI:10.1109/ACCESS.2023.3317695*
17. *J. Li and Y. Zhang. Resilient DoS Attack Detector Design for Cyber-Physical Systems // 2023 12th International Conference on Renewable Energy Research and Applications (ICRERA), Oshawa, ON, Canada, 2023, pp. 1-5. DOI:10.1109/ICRERA59003.2023.10269439*
18. *S. Roy, A. Kumar and U. P. Rao. Security Attacks and it's Countermeasures on Smart Grid: A Review // 2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3), Srinagar Garhwal, India, 2023, pp. 1–6. DOI:10.1109/IC2E357697.2023.10262686*
19. *T. Zhang and D. An. Data Integrity Attack Strategy against State Estimation Results of Distributed Power System // 2023 5th Asia Energy and Electrical Engineering Symposium (AEEES), Chengdu, China, 2023, pp. 1146-1151. DOI:10.1109/AEEES56888.2023.10114340*

20. *S. Vahidi, M. Ghafouri, M. Au, M. Kassouf, A. Mohammadi and M. Debbabi. Security of Wide-Area Monitoring, Protection, and Control (WAMPAC) Systems of the Smart Grid: A Survey on Challenges and Opportunities // in IEEE Communications Surveys & Tutorials. 2023, vol. 25, no. 2, pp. 1294–1335. DOI:10.1109/COMST.2023.3251899.*
21. *G. B. Gaggero, D. Piserà, P. Girdinio, F. Silvestro and M. Marchese. Novel Cybersecurity Issues in Smart Energy Communities // 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp. 1–6. DOI:10.1109/ICAISC56366.2023.10085312*
22. *J. Kim, S. Bhela, J. Anderson and G. Zussman. Identification of Intraday False Data Injection Attack on DER Dispatch Signals // 2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Singapore, Singapore, 2022, pp. 40–46. DOI:10.1109/SmartGridComm52983.2022.9960974*
23. *A. Huseinovic, Y. Korkmaz, H. Bisgin, S. Mrdović and S. Uludag. PMU Spoof Detection via Image Classification Methodology against Repeated Value Attacks by using Deep Learning // 2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, 2022, pp. 1–6. DOI:10.1109/ICAT54566.2022.9811128*
24. *M. D. Roig Greidanus, S. K. Mazumder and N. Gajanur. Identification of a Delay Attack in the Secondary Control of Grid-Tied Inverter Systems // 2021 IEEE 12th International Symposium on Power Electronics for Distributed Generation Systems (PEDG), Chicago, IL, USA, 2021, pp. 1–6. DOI:10.1109/PEDG51384.2021.9494253*
25. *K. P. Swain, A. Tiwari, A. Sharma, S. Chakrabarti and A. Karkare. Comprehensive Demonstration of Man-in-the-Middle Attack in PDC and PMU Network // 2022 22nd National Power Systems Conference (NPSC), New Delhi, India, 2022, pp. 213-217. DOI:10.1109/NPSC57038.2022.10069874*
26. *M. Z. Gunduz, R. Das. A comparison of cyber-security oriented testbeds for IoTbased smart grids // in: 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1–6. DOI:10.1109/ISDFS.2018.8355329*
27. *Z. E. Mrabet, N. Kaabouch, H. E. Ghazi, H. E. Ghazi. Cyber-security in smart grid: survey and challenges // Comput. Electr. Eng. 2018, 67, pp. 469–482. DOI:10.1016/j. compeleceng.2018.01.015*
28. *M. S. Al-kahtani, L. Karim. A survey on attacks and defense mechanisms in smart grids // Int. J. Comput. Eng. Inform. Technol. 2019, 11 (5), 7.*
29. *Илюшин П. В., Вольный В. С. Обзор структур микросетей низкого напряжения с распределенными источниками энергии // Релейная защита и автоматизация. 2023, № 1(50), с. 68–80.*
30. *Гурина Л. А., Томин Н. В. Разработка комплексного подхода к обеспечению кибербезопасности взаимосвязанных информационных систем при интеллектуальном управлении сообществом микросетей // Вопросы кибербезопасности. 2023, № 4(56), с. 94–104. DOI: 10.21681/2311-3456-2023-4-94-104*
31. *Колосок И. Н., Гурина Л. А. Оценка рисков управления киберфизической ЭЭС на основе теории нечетких множеств. Методические вопросы исследования надежности больших систем энергетики. В 2-х книгах. 2019, с. 238–247.*

МОДЕЛИРОВАНИЕ УСТОЙЧИВОСТИ КРИТИЧЕСКОЙ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ НА ОСНОВЕ ИЕРАРХИЧЕСКИХ ГИПЕРСЕТЕЙ И СЕТЕЙ ПЕТРИ

Бочков М. В.[1], Васинев Д. А.[2],

DOI: 10.21681/2311-3456-2024-1-108-115

Цель исследования: *моделирование объектов критической информационной инфраструктуры (КИИ) на основе математического аппарата гиперсетей и сетей Петри.*

Методы исследования: *математические методы теории систем и системного анализа методы теории графов, методы имитационного моделирования.*

Результат исследования: *предлагаемый способ построения математических моделей позволяет разработать параметрический точные имитационные модели для исследования свойств защищенности и устойчивости объектов КИИ, моделировать воздействия на них компьютерных атак (КА). В частности, предлагаемый способ имитационного моделирования позволяет учитывать конфигурационные и коммуникационные особенности построения и функционирования, динамику воздействия нарушителя на объекты КИИ, существующую политику безопасности, моделирование функциональных и структурных свойств устойчивости, исследования степени влияния этих элементов на защищенность объекта КИИ. Это обеспечивает возможность осуществлять оценку защищенности, обеспечение ИБ объектов КИИ с учетом конфигурационных и коммуникационных параметров объекта КИИ, уменьшить зависимость от экспертных оценок.*

Научная новизна: *заключается в развитии теории информационной безопасности в области оценки защищенности с учетом устойчивости и живучести объектов КИИ на основе математического аппарата иерархический гиперсетей, сетей Петри.*

Практическая ценность *заключается в получении параметрически точных моделей объекта КИИ. Возможности получения оценок защищенности на основании коммуникационных, инфраструктурных параметров самого объекта. Моделировании известных воздействий из банка данных угроз безопасности для проверки политики безопасности объекта КИИ в полученной модели. Моделировании воздействия на объект КИИ неизвестных ранее угроз.*

Ключевые слова: *информационная безопасность, коммуникационная инфраструктура, конфигурационная инфраструктура, моделирование математическое, моделирование имитационное, оценка защищенности, устойчивость, протокольные блоки данных.*

MODELING THE STABILITY OF CRITICAL INFORMATION INFRASTRUCTURE BASED ON HIERARCHICAL HYPERNETS AND PETRI NETS

Bochkov M. V.[3], Vasinev D. A.[4]

Research objective: *modeling of critical information infrastructure (CII) objects based on the mathematical apparatus of hypernets and Petri nets. The proposed method of building mathematical models allows to develop parametric accurate simulation models to study the properties of security and stability of CII objects, to simulate the impact of computer attacks (CA) on them.*

Research methods: *mathematical methods of systems theory and systems analysis methods of graph theory, methods of simulation modeling.*

Research result: *the proposed method of simulation modeling allows to take into account the configuration and communication features of construction and operation, the dynamics of the impact of the intruder on CII*

1 Бочков Максим Вадимович, доктор технических наук, профессор, ЧОУ ДПО «Центр предпринимательских рисков», г. Санкт-Петербург, Россия. E-mail: mvboch@cprspb.ru

2 Васинев Дмитрий Александрович, кандидат технических наук, сотрудник Академии ФСО России, г. Орёл, Россия. E mail: vda33@academ.msk.rsnet.ru

3 Maxim V. Bochkov, Dr. Sc., Professor, Center for entrepreneurial risks, Saint Petersburg, Russia. E mail: mvboch@yandex.ru

4 Dmitriy A. Vasinev, Ph.D., employee Academy FSO Russia, Orel, Russia. E mail: vda33@academ.msk.rsnet.ru

objects, the existing security policy, modeling of functional and structural properties of stability, research into the degree of influence of these elements on the security of the CII object. This makes it possible to assess the security, to ensure the IS of CII objects taking into account the configuration and communication parameters of the CII object, to reduce the dependence on expert assessments.

Keywords: *information security, communication infrastructure, configuration infrastructure, mathematical modeling, simulation modeling, hypernets, security assessment, stability, protocol data blocks.*

Введение

Актуальность вопросов обеспечения информационной безопасности для информационных систем (ИС), информационно-телекоммуникационных сетей (ИТС), автоматизированных систем управления (АСУТП) критических информационных инфраструктур (КИИ), функционирующих в критически важных отраслях деятельности государства в медицине, образовании, промышленности, энергетике поясняется отраслевой принадлежностью объектов атак, что говорит о продолжающемся информационном противоборстве. Среди прочих, целью нарушителя являются объекты КИИ. При этом уровень деструктивных действий нарушителя на коммуникационную инфраструктуру говорит о сетевых угрозах преимущественно высокого и критического уровней воздействия нарушителя, проявляющихся в атаках на КИИ [5,6,7].

В качестве составных элементов КИИ выступают распределенные фрагменты сетей, центры обработки данных (ЦОД), автоматизированные системы управления (АСУТП) объединенные в единую распределенную ИТС организации. Пример обобщённого представления распределенной КИИ представлен на рис. 1. Существующие особенности построения коммуникационной инфраструктуры технологически достаточно разнообразны, однако общими требованиями являются: применение технологий виртуальных частных сетей (VPN), резервирования, отказоустойчивости, обеспечение устойчивости в условиях воздействия компьютерных атак (КА)[8,9]. Кроме того, современные условия функционирования технических систем предполагают применение отечественного коммуникационного оборудования, средств защиты для проектирования новых и импортозамещения существующих фрагментов КИИ. В этих условиях исследования в области оценки защищенности и устойчивости КИИ в условиях воздействия на нее КА является актуальной задачей.

Воздействие нарушителя на распределенную ИТС, обусловлено инфраструктурными, коммуникационными особенностями организации каналов

5 РосТелеком. Аналитический отчет об атаках на онлайн ресурсы компании за 2022г. [сайт]. URL:https:// https://rt-solar.ru/upload/iblock/34a/5w4h9o57axovdbv3ng7givrz271ykir3/Ataki-na-onlayn_resursy-rossiyskikh-kompaniy-v-2022-godu.pdf.

6 ТрансТелеКом. Аналитический отчет по сервису «Защита от DDoS-атак» 1 квартал 2023 [сайт]. URL: https://ttk.ru/upload/doc/business/ddos_1_2023.pdf.

7 Бюллетени НКЦКИ: новые уязвимости ПО [сайт]. URL: https://safe-surf.ru/specialists/bulletins-nkcki/.

8 Запечников, С. В. Основы построения виртуальных частных сетей: учебное пособие для вузов/ Запечников, С.В., Милославская, Н. Г., Толстой, А. И. – 2-е изд. Москва: Горячая линия-Телеком, 2011, – 249. – ISBN 5-85582-119

9 Захватов, М. А. Построение виртуальных частных сетей на базе технологии MPLS / М. А. Захватов. – Москва: изд-во Cisco Systems, 2001 г.

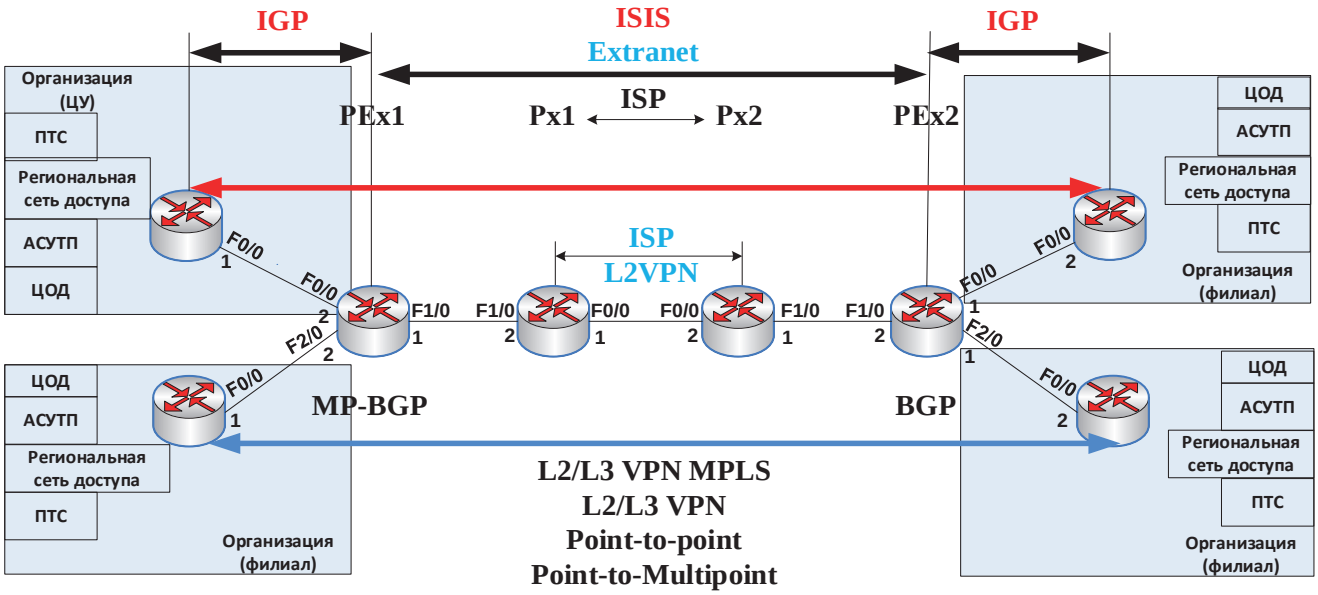


Рис. 1. Формирование распределенной инфраструктуры для объектов ИС, АСУТП, ИТС КИИ

связи, предлагаемых оператором связи, на основе которого осуществляется организация взаимодействия между распределенными филиалами телекоммуникационных объектов КИИ, представлено на рис. 2. Сетевые, транспортные и управляющие протоколы, которые применяются в коммуникационных инфраструктурах для передачи данных, управления, такие как (Ethernet, ICMP, IP, TCP, SNMP, Modbus, MMS, Goose). Для выделенных протоколов помимо иерархических – коммуникационных особенностей, можно выделить конфигурационные компоненты формирования инфраструктур, которые также могут быть причиной снижения защищенности объекта – в связи с воздействием нарушителя, или неквалифицированных действий персонала в распределенных фрагментах ИТС.

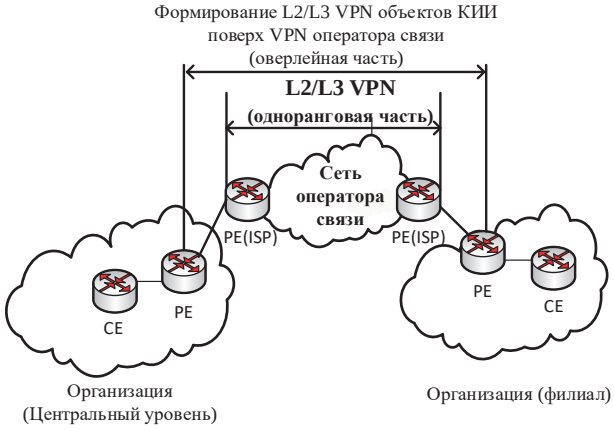


Рис. 2. Формирование распределенной инфраструктуры для ИС, АСУТП, ИТС

В настоящее время при обеспечении информационной безопасности (ИБ) объектов КИИ наряду со свойствами целостности, доступности, конфиденциальности, формируется понятие устойчивости КИ. Так, например, в нормативных документах[10,11,12] и известных исследованиях в области ИБ ряд авторов рассматривает свойство устойчивости объектов коммуникационной инфраструктуры от компьютерных атак [1–3] как свойство защищенности объекта, связанное с его способностью противостоять КА.

Решение задачи устойчивости функционирования КИИ в работах [1–3] связывается с возможностями противостоять компьютерным атакам методом резервирования на структурном – физическом уровне, как например в [4], на основе вероятностного расчета риска и нечетких множеств [5], вероятностных и Марковских методов оценки защищенности в работе [6]. Анализ вариантов оценок киберустойчивости объектов КИИ, представленные в работах [7–9], показывает, что в оценках киберустойчивости авторы применяют вероятностные методы, теорию нечетких множеств [7], основываются на экспертном методе при формировании алгоритмов оценки киберустойчивости в работе [8]. Авторы [9] предлагают применять описательные и концептуальные модели динамики обеспечения киберустойчивости объекта КИИ. Таким образом, при оценках киберустойчивости выделенные методы теории рисков, теории вероятностей, нечетких множеств, экспертные методы не учитывают иерархические и параметрические особенности построения и функционирования объектов КИИ, а также иерархические и параметрические особенности воздействия нарушителя. Как при оценке защищенности, так и при оценках киберустойчивости применяют обобщенные или абстрактные показатели, характеризующие защищённость, осуществляют свертку таких показателей в обобщенный показатель, характеризующий защищенность объекта. В рамках исследования делается предположение о возможностях получения оценок защищенности на основе иерархических особенностей построения объекта КИИ, параметров ее конфигурации, параметров воздействия нарушителя. На основе выделенных параметрических особенностей предлагается способ обеспечения устойчивости объектов КИИ при противодействии КА в логических каналах методом динамического изменения характеристик функционирования, что соответствует функциональному варианту обеспечения устойчивости. Примером логического резервирования могут быть параметры самого логического канала, типы применяемых виртуальных частных сетей (VPN), топология соединения, маршрутная информация, скорость передачи, качество обслуживания. Все это связано с технологическими особенностями построения, применимыми технологиями, иерархическими особенностями построения КИИ, вариант которой представлен на рис. 3.

Технологические особенности связаны с применением различных вариантов туннелирования (L2, L3 VPN), при формировании распределенной ИТС, а также с применением как физически зарезервированных каналов, так и логического резервирования на основе следующих технологий и протоколов (Rapid spanning tree protocol (RSTP), Virtual router redundancy protocol (VRRP), Bidirectional forwarding detection (BFD), Routing, MPLS Fast reroute FastRR) [6, 7].

10 Основные направления государственной политики в области обеспечения безопасности автоматизированных систем управления производственными и технологическими процессами критически важных объектов инфраструктуры Российской Федерации (утв. Президентом РФ 03 февраля 2012 г., № 803). Режим доступа: https://fstec.ru/component/attachments/download/1906.

11 О безопасности Критической информационной инфраструктуры Российской Федерации: Федеральный закон ред. от. 19.07.2017г. №187 // ФСТЭК: [сайт]. – URL: https://fstec.ru/component/attachments/download/1906/.

12 Об утверждении Требований по обеспечению безопасности значимых объектов критической информационной инфраструктуры Российской Федерации: Приказ ФСТЭК России № 239 от 25.12.2017 // ФСТЭК: [сайт]. – URL: https://fstec.ru/dokumenty/vse-okumenty/prikazy/prikaz-fstek-rossii-ot-25-dekabrya-2017-g-n-239/.

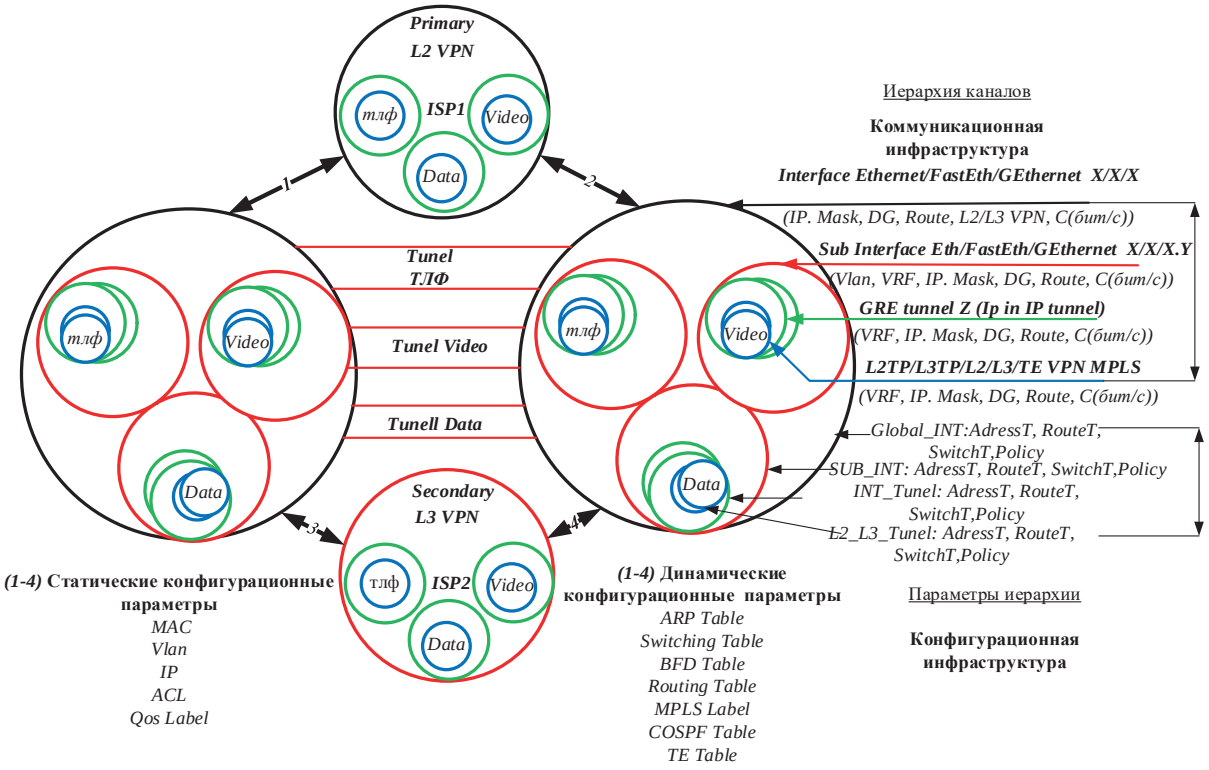


Рис.3. Вариант иерархической, вложенной коммуникационной инфраструктуры для моделирования распределенной инфраструктуры ИС, АСУТП, ИТС, объектов КИИ

Очевидно, что логическая структура каналов связи для КИИ имеет иерархическую особенность построения, обусловленную применением коммуникационных и конфигурационных параметров в КИИ рассматриваемых объектов (ИС, АСУТП, ИТС), функционирующих в единой распределенной сети организации.

Для оценки защищенности объектов (ИС, АСУТП, ИТС) а также исследования свойств устойчивости, с учетом иерархических особенностей КИИ предлагается применять математические модели основаны на теории гиперграфов[13] и [12]. При этом отличительная особенность предлагаемого решения на основе гиперграфов является учет не только иерархических особенностей построения объектов КИИ, но и их конфигурационных и коммуникационных особенностей функционирования, а также воздействий нарушителя как на логическую (коммуникационную и конфигурационную), так и на физическую составляющую объекта КИИ.

Моделирование объектов критической информационной инфраструктуры

Моделированию устойчивости объектов КИИ посвящены работы [10,11], однако не учитываются параметрические особенности моделирования протокольных единиц данных, иерархические особенности их построения и функционирования. Математическая модель на основе иерархических гиперграфов, позволяет наиболее полно отразить характеристики моделируемого объекта, связанные с иерархичностью и вложенностью протекающих процессов[14,15] и [12–14]. Развитие теории гиперграфов в области информационной безопасности связано с оценкой защищенности АСУТП на основе наиболее вероятной угрозы [15]. Предлагаемое решение заключается в учете конфигурационных и коммуникационных возможностей, а также конфликтности, связанной с воздействием нарушителя на выделенные структурные элементы КИИ в гиперграфе.

Математическая модель КИИ, учитывающая иерархические, вложенные инфраструктурные, конфигурационные компоненты, на основе теории s-гиперсетей представлена выражением (1)

$$H = (G_0, G_1, \ldots, G_m, G_{ИБ}, G_V), \quad (1)$$

где: G_0 – граф первичной физическая топология; $G_{1\ldots m}$ – графы инфраструктурных компонентов (технологий, протоколов, виртуальных туннелей), конфигураций;

13 Зыков А. А. Гиперграфы / Успехи математических наук, 11974. Т.9, выпуск 6, 89–154 // Общероссийский математический портал Math-Net.Ru [сайт] – URL:https://www.mathnet.ru/links/6ebfd77a48733caf850ac105bc7eaac6/rm4449.pdf.

14 Конин М. В. Применение S–гиперсетей для автоматизированного проектирования инженерной инфраструктуры предприятия /М. В. Конин, Э. Ю. Лепнер, Г. В. Попков / Информационные технологии в системах автоматизации. 2013. – №5 (24). [сайт]. – URL: https://www.elibrary.ru/download/elibrary_28906854_84188973.pdf

15 Лепнер Э. Ю. Разработка операций над S-гиперсетями: дис. на соиск. учен. магистра / Лепнер Эдуард Юрьевич: Новосибирский национальный исследовательский государственный университет. – Новосибирск – 2013: [сайт]. – URL: https://www.elibrary.ru/download/elibrary_28906854_84188973.pdf.

$G_{ИБ}$ – гиперграф политик информационной безопасности; G_V – гиперграф воздействия нарушителя.

На основании анализа работ по теории s-гиперсетей [12,13] и [12–15], а также руководствуясь возможностями математического моделирования на основе теории иерархических s-гиперсетей установлено, что исследование на таких моделях динамики функционирования объекта КИИ, а также конфигурация s-гиперсетей и исследования защищенности, моделирование динамики воздействия нарушителя имеет ограничения, связанные со сложностью задания выделенных объектов динамическими матрицами смежности или инцидентности, а также динамического преобразования исходного гиперграфа к необходимому виду.

Для устранения выделенных недостатков на данном этапе работы осуществлялось применение функционального аппарата раскрашенных вложенных сетей Петри, позволяющего моделировать и исследовать динамику функционирования КИИ (изменчивость гиперграфа в различных условиях). Примерами динамики изменения гиперграфа могут быть воздействия нарушителя, связанные с изменением коммуникационной и конфигурационной компонент гиперграфа, в условиях функционирования как объекта КИИ, так средств обеспечения ИБ (существующих политик ИБ), которые учтены в модели, представленной выражением (2)

$$S = (\{P_Д, P_{ИБ}, P_V\},\{T_Д, T_{ИБ}, T_V\},\{E_Д, E_{ИБ}, E_V\}, M_0), \quad (2)$$

где $P_Д$ – конечное множество допустимых позиций КИИ УК; $P_{ИБ}$ – конечное множество позиций политик ИБ; P_V – конечное множество воздействий нарушителя; $T_Д$ – конечное множество допустимых переходов (событий); $T_{ИБ}$ – конечное множество переходов (событий) политик информационной безопасности; T_V – конечное множество воздействий нарушителя; $E_Д$ – конечное множество дуг допустимых переходов (событий); $E_{ИБ}$ – конечное множество дуг событий политик ИБ; E_V – конечное множество дуг событий воздействия нарушителя; M_0 – начальное состояние сети.

Пример моделирования объектов коммуникационной инфраструктуры иерархическими s-гиперсетями представлен на рис. 4–8. Диаграмма вложений протоколов формирующих КИИ гиперграфов в s-гиперсети для ИТС КИИ представлена на рис. 4. Гиперграф первичной сети рис. 4, формирует первичную топологию ИТС КИИ с ее основными коммуникационными элементами – узлами, физическими каналами связи – ребрами гиперграфа.

Графы второго и последующих уровней вложения (рис. 5–8) определяются вершинами и рёбрами, участвующими во взаимодействии протоколов соответствующего уровня.

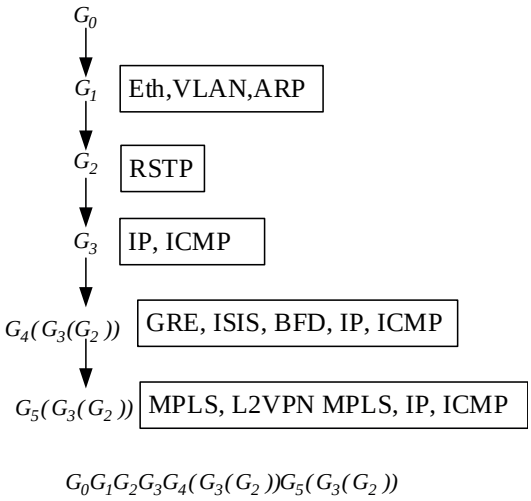


Рис. 4. Диаграмма вложений гиперграфов в s-гиперсети

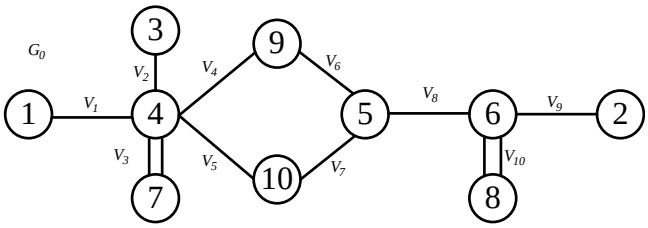


Рис. 5. Топология первичного графа G0 для ИТС объекта КИИ

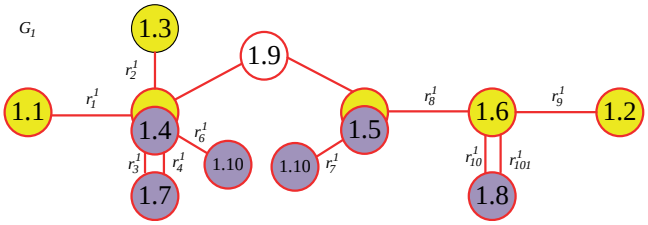


Рис. 6. Топология графа G1 (VLAN, ARP) для КИИ

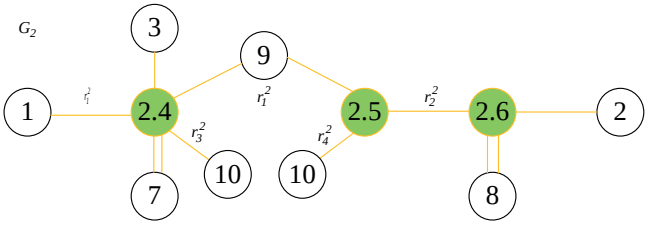


Рис. 7. Топология графа G2 (RSTP) для КИИ

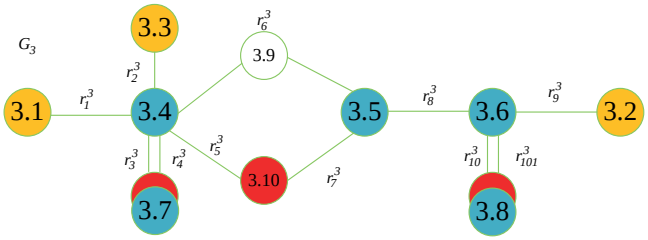


Рис. 8. Топология графа G3 (IP, ICMP) для КИИ

На рис. 4 представлена первичная топология – G0 объекта КИИ, узлы: 1, 2, 3 – оконечное оборудование (автоматизированные системы, АСУТП);

4, 5, 6 – коммутационное оборудование класса L2; 7, 8 маршрутизирующее оборудование L3; 9, 10 – оборудование L2, L3; оператора связи.

Вершины гиперграфа в сети Петри – это протокольные блоки данных, функционирующие на различных уровнях модели взаимодействия открытых систем (рис. 9–10), обладающие строго формализованными функциональными свойствами. При этом основная причина выбора имитационного моделирования заключается в сложности процесса формирования протокольного блока данных, взаимосвязи его с другими иерархическими элементами, размерности параметрического пространства состояний, что усложняет на данном этапе работы применение s-гиперсетей. Переход от гиперграфового подхода к формированию узлов графа в виде матриц смежности или инцидентности – к заданию вершин гиперграфа протокольными блоками данных в сети Петри, позволяет устранить выявленные на данном этапе работы недостатки в области масштабирования моделей, динамики смены состояний параллельных и асинхронных процессов, которые характерны для ИС, АСУТП, ИТС объектов КИИ.

Имитационное моделирование позволяет разработать универсальные способы построения имитационных протокольных блоков данных, для различных типов протоколов, учесть коммуникационные и конфигурационные особенности их функционирования (рис. 9–10).

Примеры моделирования вариантов конструкций протокольных блоков данных и особенностей их объединения, представлены на рис. 11–14.

Способ формирование кадров различной структуры параметрически полно отражающих существующие коммуникационные особенности объектов КИИ в процессе отправки, получения, обработки, представлен на рис. 11.

Учет полноты моделируемых параметров, позволяет сформировать многообразие пространства состояний протокольных блоков данных, что с одной стороны усложняет решение задачи оценки защищенности, а с другой позволяет задать точнее ограничения на пространстве состояний, учесть динамику воздействия нарушителя, проверять работоспособность политики безопасности.

Пример объединения различных типов кадров (кадр Ethernet, arp – запрос, arp – ответ) в едином канале обработки данных представлен на рис. 12, а демультиплексирования кадров – выбора из единого канала кадров заданного формата для дальнейшей их обработки на рис. 13.

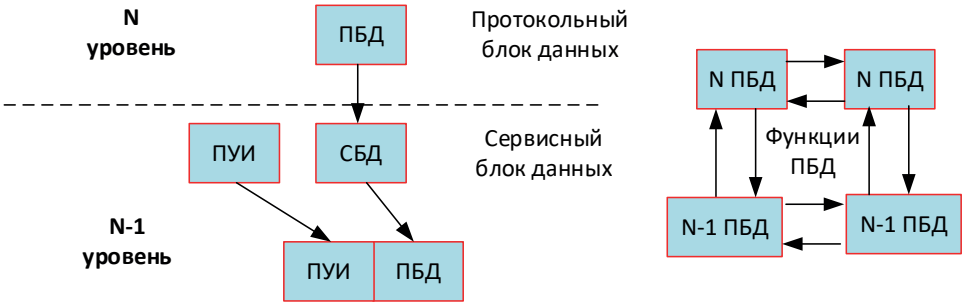


Рис. 9. Методы, способы взаимодействия ПБД в соответствии с моделью OSI (7498), X.200, ГОСТ Р ИСО/МЭК 7498-1-99

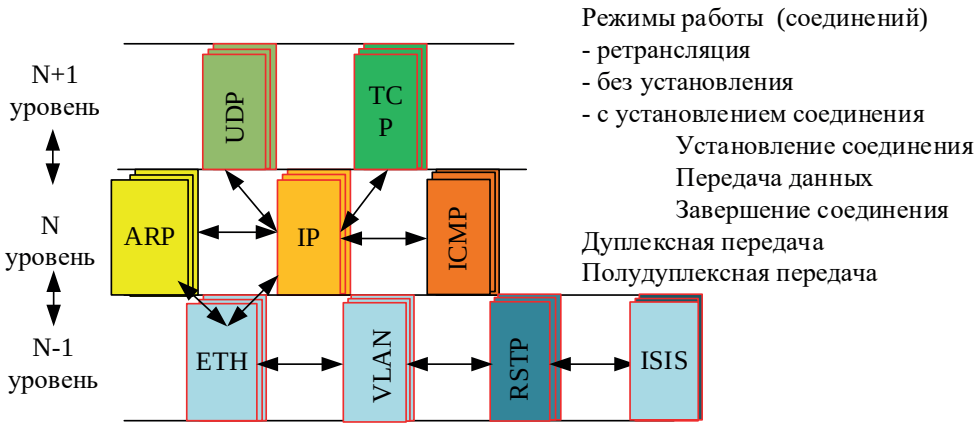


Рис.10. Вариант взаимодействия объектов различных уровней в соответствии с моделью OSI (7498), X.200, ГОСТ Р ИСО/МЭК 7498-1-99

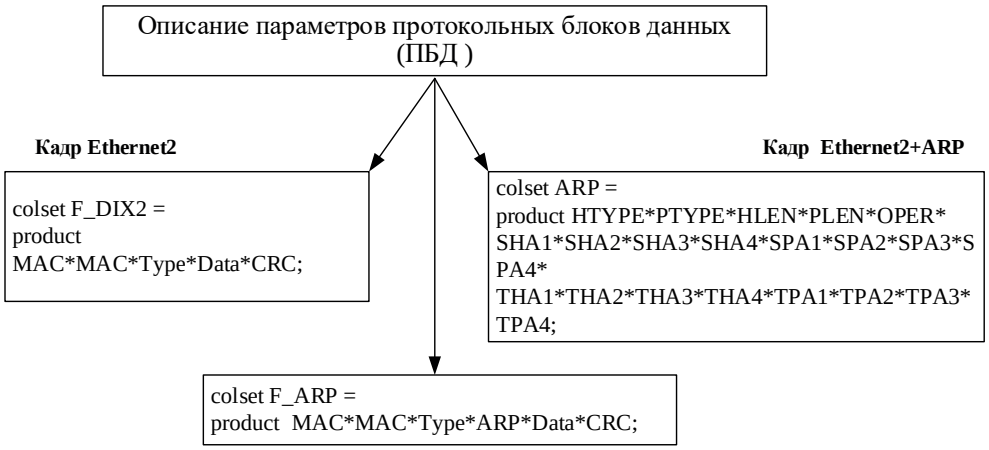


Рис. 11. Формирование иерархии цветов (типов) для кадра Ethernet, протокола ARP

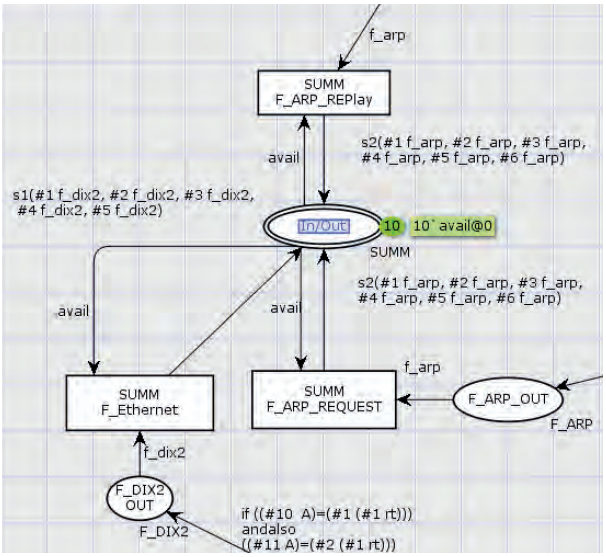


Рис. 12. Мультиплексирование объединение кадров с различным содержимым классов (F_DIX2, f_arp – запрос, f_arp – ответ) в структуры S1, S2 для передачи в канал

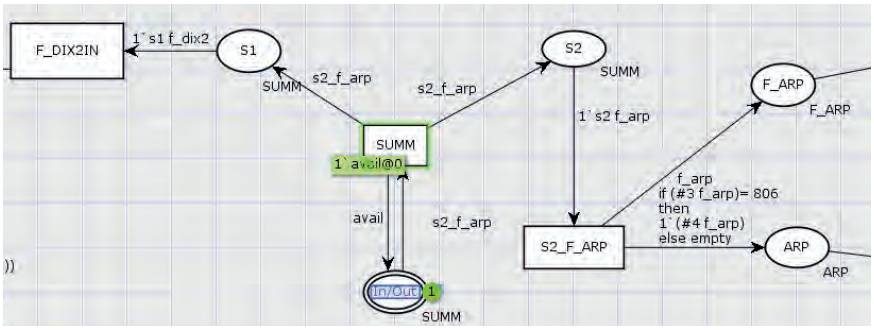


Рис. 13. Расщепление из единого потока S1, S2, кадров Ethernet с данными (F_DIX2), кадров протокола ARP (f_arp), пакетов протокола ARP (F_DIX2, f_arp – запрос, f_arp – ответ) в структуры S1, S2

Пример моделирования единой иерархической коммуникационной инфраструктуры на примере взаимодействия Ethernet, ARP представлен на рис. 14.

Заключение

Таким образом, моделирование фрагментов КИИ (ИС, АСУТП, ИТС) на основе математического аппарата иерархических s-гиперсетей и сетей Петри позволяет расширить прикладной аспект теории информационной безопасности в направлении моделирования и оценки защищенности объектов КИИ с учетом ее устойчивости в условиях воздействия КА. Моделирование на основе сетей Петри позволяет исследовать влияние протокольных особенностей построения объектов КИИ (ИС, АСУТП, ИТС) на свойства

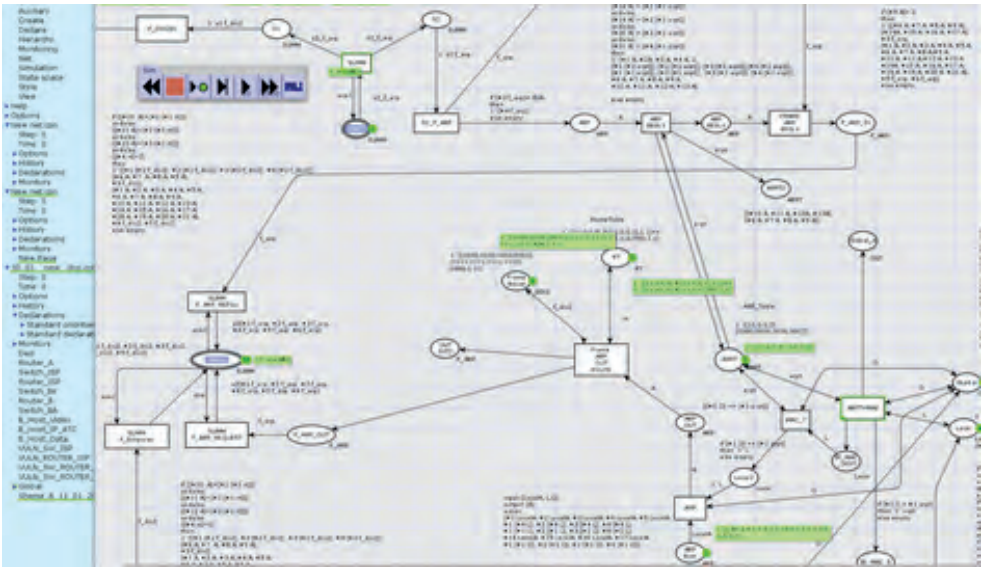


Рис. 14. Пример представления протокола ARP в имитационной модели

устойчивости и доступности объектов КИИ, и оценивать на основе этого их защищенность. Формирование параметрически точных моделей КИИ как аналитических, так и имитационных, позволяет строить цифровые двойники объектов коммуникационной инфраструктуры и в динамике исследовать функционирование такого объекта с учетом изменения конфигурации, воздействия нарушителя, формирования физических или логических резервных направлений связи. Полученные результаты позволяют получать в том числе и количественные показатели оценки защищенности объектов КИИ в условиях воздействия нарушителя и исследовать влияние на них различных типов компьютерных атак.

Литература

1. *Зегжда Д. П. Кибербезопасность цифровой индустрии. Теория и практика функциональной устойчивости к кибератакам / Под редакцией профессора РАН, доктора технических наук Д.П. Зегжды. – Москва: Горячая линия – Телеком. 2023. – 500с. – ISBN 978-5-9912-0827-7.*
2. *Петренко С. А. Киберустойчивость цифровой индустрии 4.0: научная монография / С. А.Петренко. – Санкт-Петербург: Издательский Дом «Афина», 2020, – 256 с.*
3. *Петренко С. А. Управление киберустойчивостью. Постановка задачи // Защита информации. Инсайд. 2019. № 3(87). С. 16–24.*
4. *Штыркина А. А. Обеспечение устойчивости киберфизических систем на основе теории графов. Проблемы информационной безопасности // Компьютерные системы. 2021. № 2. С. 145–150.*
5. *Колосок И. Н., Гурина Л. А. Оценка показателей киберустойчивости систем сбора и обработки информации в ЭЭС на основе полумарковских моделей // Вопросы кибербезопасности, 2021, № 6(46), С. 2-11. DOI: 10.21681/2311-3456-2021-6-2-11*
6. *Гурина Л. А. Повышение киберустойчивости SCADA и WAMS при кибератаках на информационно-коммуникационную подсистему ЭЭС // Вопросы кибербезопасности. 2022. №2(48). С.18–26. DOI: 10.21681/2311-3456-2022-2-18-26*
7. *Гурина Л. А. Оценка киберустойчивости системы оперативно-диспетчерского управления ЭЭС // Вопросы кибербезопасности, 2022. № 3(48), С.18–26. DOI: 10.21681/2311-3456-2022-3-23-31*
8. *Чиркова Н. Е. Анализ существующих подходов к оценке киберустойчивости гетерогенных систем // Сборник материалов Международной научно-практической конференции: Техника и безопасность объектов уголовно-исполнительной системы Иваново. 2022. С. 408–410.*
9. *Бобров В. Н., Захарченко Р. И., Бухаров Е. О., Калач А. В. Системный анализ и обоснование выбора моделей обеспечения киберустойчивого функционирования объектов критической информационной инфраструктуры //Вестник Воронежского института ФСИН России. 2019. № 4. С. 31–43.*
10. *Минаев М. В., Бондарь К. М., Дунин В. С. Моделирование киберустойчивости информационной инфраструктуры МВД России // Криминологический журнал. 2021. № 3. С. 123–128.*
11. *Осипенко А. А., Чирушкин К. А., Скоробогатов С. Ю., Жданова И. М., Корчевной П. П. Моделирование компьютерных атак на программно-конфигурируемые сети на основе преобразования стохастических сетей //Известия Тульского государственного университета. Технические науки. 2023. № 2. С. 274–281.*
12. *Ванг Л., Егорова Л. К., Мокряков А. В., Развитие теории Гиперграфов // Известия РАН. Теория и системы управления. 2018. №1. С. 111–116. DOI: 10.7868/S00023388180110.*
13. *Величко В. В. Модели и методы повышения живучести современных систем связи / В. В. Величко, Г. В. Попков, В. К. Попков. – Москва: Горячая линия – Телеком, 2017. – 270 с. ISBN 978-5-94876-090-2.*
14. *Попков Г. В. Математические основы моделирования сетей связи / В. В. Величко, Г. В. Попков, В. К. Попков. – Москва: Горячая линия – Телеком, 2018.–182 с. ISBN 978-5-9912-0266-4.*
15. *Barrere M., Hankin C., Nicolaou N. // Journal of Information Security and Application. 2020. №52. DOI: 10.1016/jisa.2020.102471 [сайт]. – URL: https://www.sciencedirect.com/science/article/pii/S2214212619311342?via%3Dihub(дата обращения 10.11.2023).*

ОЦЕНКА РИСКОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ ТЕХНОЛОГИЧЕСКИМ ПРОЦЕССОМ

Иваненко В. Г.[1], Иванова Н. Д.[2]

DOI: 10.21681/2311-3456-2024-1-116-123

Цель работы: *формирование методики количественной и качественной оценки рисков информационной безопасности автоматизированных систем управления технологическим процессом как объектов критической информационной инфраструктуры и предложений об ее внедрении в проводимый процесс категорирования объектов критической информационной инфраструктуры с целью адаптации базового набора мер защиты.*

Метод исследования: *анализ существующих подходов к оценке рисков информационной безопасности. Анализ отечественных и зарубежных нормативно-правовых и методических документов на предмет применимости для оценки рисков информационной безопасности автоматизированных систем управления технологическим процессом. Построение блок-схем процессов оценки рисков.*

Результаты: *в исследовании обоснована необходимость проведения оценки рисков информационной безопасности автоматизированных систем управления технологическим процессом с целью адаптации базового набора мер защиты. Проведен анализ методов количественной и качественной оценки рисков информационной безопасности, определен смешанный подход к оценке рисков как компромиссный между ними. На основании национальных и международных нормативно-методических документов и практики обеспечения информационной безопасности были определены факторы и характеристики рисков информационной безопасности, а также возможность их количественной оценки. Сформированы предложения к алгоритму количественной и качественной оценки рисков информационной безопасности автоматизированных систем управления технологическим процессом и к его внедрению в проводимый процесс категорирования объектов критической информационной инфраструктуры. Составлены блок-схема соответствующих процессов.*

Практическая ценность: *практическая ценность работы заключается в предложении метода оценки рисков, согласованного с существующей практикой обеспечения информационной безопасности автоматизированных систем управления технологическим процессом, методами управления рисками информационной безопасности и требованиями регулирующих органов. Результаты проведенного анализа и выработанные рекомендации по адаптации базового набора мер защиты могут быть используемы для повышения информационной безопасности автоматизированных систем управления технологическим процессом.*

Ключевые слова: *угрозы информационной безопасности, уязвимости, базовый набор мер защиты, CVSS, количественные, качественные, смешанные (гибридные) методы оценки рисков.*

INFORMATION SECURITY RISK ASSESSMENT OF INDUSTRIAL CONTROL SYSTEMS

Ivanenko V. G.[3], Ivanova N. D.[4]

Purpose: *development of a methodology for quantitative and qualitative assessment of information security risks of industrial control systems as objects of critical information infrastructure and development of proposals*

1 Иваненко Виталий Григорьевич, доктор технических наук, профессор Института Интеллектуальных Кибернетических Систем (ИИКС) Национального исследовательского ядерного университета «МИФИ», г. Москва, Россия. E-mail: VGIvanenko@mephi.ru

2 Иванова Нина Дмитриевна, аспирант кафедры «Управление и защита информации» Российского университета транспорта (МИИТ), г. Москва, Россия. E-mail: IvanovaND.Nina@yandex.ru, ORCID 0000-0001-5942-8050

3 Vitaly G. Ivanenko, Dr.Sc., Associate Professor of the Institute of Intelligent Cybernetic Systems of the National Research Nuclear University «MEPhI», Moscow, Russia. E-mail: VGIvanenko@mephi.ru

4 Nina D. Ivanova, assistant of the Department of Management and Information Security, Russian University of Transport (MIIT), Moscow, Russia. Email: IvanovaND.Nina@yandex.ru, ORCID 0000-0001-5942-8050

for its implementation in addition to the process of categorizing objects of critical information infrastructure in order to adapt the basic set of protection measures.

Research method: *analysis of existing approaches to assessing information security risks. Analysis of national and international regulatory and methodological documents for applicability for assessing the information security risks of industrial control systems. Drawing up flowcharts of risk assessment processes.*

Results: *the study substantiates the need to conduct an assessment of the information security risks of industrial control systems in order to adapt a basic set of protection measures. An analysis of methods for quantitative and qualitative assessment of information security risks was carried out, and a hybrid approach to risk assessment was determined as a compromise between them. Based on national and international regulatory and methodological documents and information security practices, the factors and characteristics of information security risks of industrial control systems were identified, as well as the possibility of their quantitative assessment. Proposals have been formulated for an algorithm for quantitative and qualitative risk assessment of industrial control systems and for its implementation in addition to the process of categorizing objects of critical information infrastructure. Flowcharts of the relevant processes have been drawn up.*

Practical value: *the practical value of the work lies in the proposal of a risk assessment method consistent with the existing practice of ensuring the information security of industrial control systems, information security risk management methods and the requirements of regulatory authorities. The results of the analysis and recommendations developed for adapting a basic set of protection measures can be used to improve the information security of industrial control systems.*

Keywords: *information security threats, vulnerabilities, basic set of protection measures, CVSS, qualitative, qualitative, hybrid risk assessment methods.*

Введение

До сравнительно недавнего времени обеспечение информационной безопасности (ИБ) не являлось приоритетной задачей для автоматизированных систем управления технологическим процессом (АСУ ТП) [1]. Безопасность ранних АСУ ТП достигалась за счет контроля физического доступа к компонентам системы – специализированным программно-аппаратным комплексам, использующим проприетарные протоколы связи.

Современные системы АСУ ТП сложны и основаны на передовых технологиях. Возрастающая сложность и модернизация, а также непрерывная работа в режиме реального времени и распределенная многокомпонентная архитектура лежат в основе роста компьютерных атак на АСУ ТП. АСУ ТП подвержены широкому спектру компьютерных атак, в том числе из-за стандартизации коммуникационных протоколов и аппаратных компонентов, растущей взаимосвязи и наследия [2].

В 2010 году компьютерный червь Stuxnet поразил иранский ядерный объект, вызвав отказ почти пятой части всех центрифуг [3]. В 2014 украинские электросети были атакованы с помощью вредоносного программного обеспечения Black Energy 3, что привело к временному обесточиванию большей части Украины [4]. В 2017 году в системах противоаварийной защиты саудовского нефтехимического предприятия была обнаружена вредоносная программа Triton/Triconex [5]; последствием успешной реализации компьютерной атаки с использованием Triton/Triconex могла стать гибель людей. В начале 2022 года группой хакеров Hackers-Arise были реализованы многочисленные нападения на объекты промышленной инфраструктуры Российской Федерации [6]. В 2010 году в Репозитории инцидентов промышленной безопасности (The Repository of Industrial Security Incidents – RISI) был зарегистрирован 161 компьютерный инцидент, причем каждый квартал добавлялось около 10 новых инцидентов. В 2013 году база данных RISI содержала уже 240 инцидентов. Начиная с января 2015 года, RISI перестал обновляться, содержа в своей базе более 300 инцидентов компьютерной безопасности. Неуклонный рост компьютерных атак на объекты промышленной автоматизации стал причиной появления соответствующих нормативных документов и требований к ИБ АСУ ТП [7].

Необходимость обеспечения ИБ АСУ ТП как объектов критической информационной инфраструктуры (КИИ) обусловлена требованиями приказов ФСТЭК России № 239[5] и 31[6]. Данные приказы содержат требования к применяемым мерам защиты информации АСУ ТП КИИ в соответствии с присвоенной категорией значимости объекта КИИ и классом защищенности АСУ. В случае пересечения требований необходимо применять наиболее строгое.

5 Приказ ФСТЭК России от 25.12.2017. № 239. URL: https://fstec.ru/normotvorcheskaya/akty/53-prikazy/868-prikaz-fstek-rossii-ot-14-marta-2014-g-n-31

6 Приказ ФСТЭК России от 14.03.2014. № 31. URL:https://fstec.ru/en/53-normotvorcheskaya/akty/prikazy/1592-prikaz-fstek-rossii-ot-25-dekabrya-2017-g-n-239

Оценка рисков безопасности АСУ ТП КИИ является основой для принятия решений, связанных с предотвращением аварий или минимизацией их негативных последствий. В соответствии с приказами ФСТЭК России базовый набор мер защиты информации (по обеспечению безопасности) должен быть адаптирован, если он не позволяет обеспечить блокирование всех угроз ИБ. Критерии для адаптации базового набора мер в приказах ФСТЭК России отсутствуют, что обуславливает необходимость разработки методики оценки рисков АСУ ТП КИИ.

Согласно [8] каждый новый разработанный нормативно-правовой акт или методический документ способствует появлению новых потенциальных рисков для защищаемой организации. Поэтому для оценки рисков ИБ АСУ ТП КИИ предлагается исследовать применимость существующих нормативно-правовых актов, государственных стандартов и методических документов. Анализ применимости этих документов и их учет в контексте управления рисками ИБ может послужить основой для разработки рекомендаций по поддержанию актуальности рисков ИБ АСУ ТП КИИ.

Целью настоящей статьи является формирование методики количественной и качественной оценки рисков ИБ (установления значения рисков ИБ) АСУ ТП КИИ как одного из этапов процесса оценки рисков ИБ, а также формирование предложений по внедрению разработанной методики в процесс категорирования объектов КИИ с целью уточнения базового набора мер защиты.

Количественные и качественные методы оценки рисков ИБ АСУ ТП КИИ

В области оценки рисков ИБ имеют широкое распространение два основных направления [9]: количественная оценка рисков (риск измеряется, например, в возможных финансовых потерях) и качественная оценка рисков (риск задается значениями лингвистической переменной).

Качественные методы оценки рисков не оперируют числовыми данными, представляя результат в виде описаний, сценариев угроз ИБ и рекомендаций. К основным недостаткам качественных методов оценки рисков можно отнести отсутствие числового представления результатов, невысокую точность и приближенный характер результатов [10]. С применением качественного подхода возможно учесть те риски, которые нельзя характеризовать количественно, с другой стороны, качественная оценка усложняет принятие точных решений по снижению рисков.

Методы количественной оценки рисков ИБ используют фактические данные, которые можно измерить математически или с помощью других вычислительных методов. Количественные методы оценки рисков учитывают только те риски, что могут быть количественно выражены (например, риски функциональных компонентов и конфигурации системы) [11]. Благодаря измеримости и воспроизводимости данных количественная оценка рисков является надежным и эффективным методом, но его недостатком является возможность игнорирования рисков, возникающих из нетехнических аспектов.

Как качественная, так и количественная оценка являются ключевыми факторами успешной деятельности по управлению рисками, и обычно их используют совместно. Например, на этапе установления контекста риска первым используют качественный подход, выявляя приоритетные риски, которые после будут уточнены с помощью количественного подхода.

Компромиссом между подходами качественной и количественной оценки может быть смешанный подход [12] к оценке рисков, также называемый гибридной оценкой рисков [13]. Смешанный подход объединяет качественный и количественный методы: например, путем перевода качественно определенного значения риска в количественный по соответствующей числовой шкале или наоборот (сопоставление значениям лингвистических переменных числовых шкал). С использованием смешанного подхода к оценке рисков сохраняются достоинства обоих методов (точность оценок, полученных из количественного метода и возможность всестороннего анализа, получаемого с использованием качественного метода оценки) и нивелируются их недостатки.

Формирование перечня факторов и характеристик рисков ИБ АСУ ТП КИИ

Под угрозой информационной безопасности следует понимать потенциальное нежелательное опасное событие, когда как риск информационной безопасности определяет степень опасности воздействия нежелательного события на систему или объект системы [14]. В определениях государственных и международных стандартов и руководств понятие риск чаще всего характеризуется как сочетание тяжести и вероятности наступления опасного события. Стандарты, определяющие риск ИБ, характеризуют его как потенциальную возможность успешно использовать уязвимость с целью создания угрозы активу, приводящей к нежелательным последствиям для организации. Следовательно, важнейшими факторами риска являются тяжесть последствий и вероятность наступления опасного события. Вероятность наступления опасного события ИБ может быть характеризована исходной защищенностью системы (уязвимостями системы и ее компонентов) и потенциалом (возможностями) нарушителя.

В результате риск ИБ определяется следующими факторами:

- величина тяжести возможных последствий от реализации опасного события;
- вероятность наступления опасного события, определяемая уязвимостями системы и ее компонентов и потенциалом нарушителя.

Согласно требованиям приказов ФСТЭК России, величину тяжести ущерба для обеспечения ИБ АСУ ТП КИИ характеризуют значения показателей критериев значимости объектов КИИ РФ и степень возможного ущерба от нарушения свойств конфиденциальности, целостности и доступности информации.

Уязвимости АСУ ТП КИИ могут быть определены с использованием стандарта Common Vulnerability Scoring System (CVSS)[7] – открытого стандарта, используемого для оценки уязвимостей, в том числе, и для уязвимостей из базы данных угроз (БДУ) ФСТЭК России. Для определения характеристик потенциала нарушителя может быть использован стандарт ГОСТ Р ИСО/МЭК 18045-2013[8], который предлагает методику определения потенциала нападения нарушителя, ориентированную на имеющиеся в системе уязвимости (и, следовательно, объекты защиты, что согласуется с определенным в [15] подходом к обеспечению ИБ АСУ ТП КИИ).

Ниже представлены факторы и характеристики рисков ИБ АСУ ТП КИИ, включая возможность их количественной оценки (табл. 1). Если количественную оценку рисков реализовать не представляется возможным, следует провести смешанную оценку рисков: качественно определенные характеристики перевести в количественные с помощью соответствующей числовой шкалы, сопоставляющей значения лингвистических переменных числовым показателям.

Большинство из предложенных характеристик определяются на основании отечественных или зарубежных методических документов, что позволяет проводить оценку рисков с использованием результатов уже проведенных исследований при их наличии.

Формирование предложений к алгоритму количественной и качественной оценки рисков ИБ АСУ ТП КИИ

Целью этапа количественной и качественной оценки рисков ИБ в процессе управления рисками ИБ является формирование количественных и качественных показателей факторов рисков для дальнейшей сравнительной оценки. Ниже приведена блок-схема процесса количественной и качественной оценки рисков ИБ АСУ ТП КИИ (рис. 1).

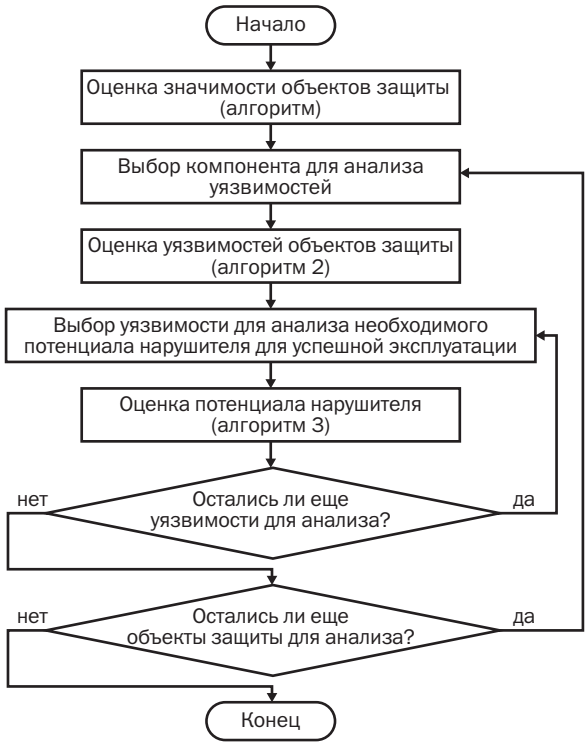


Рис. 1. Блок-схема процесса количественной и качественной оценки рисков ИБ АСУ ТП КИИ

Оценка значимости объектов защиты (рис. 2) производится на основании ранее проведенного категорирования и определения класса защищенности АСУ. В рамках анализа ущерб из-за отказов компонентов определяется тяжестью ущерба отказа соответствующей системы согласно положениям Постановления Правительства № 127[9] в части категорирования объектов КИИ, заключающимся в присвоении определенной категории значимости объектам по результатам анализа «сверху-вниз».

Рис. 2. Блок-схема оценки значимости объектов защиты АСУ ТП КИИ (алгоритм 1)

7 Common Vulnerability Scoring System version 3.1: Specification Document. – URL: https://www.first.org/cvss/specification-document/

8 ГОСТ Р ИСО/МЭК 18045–2013 (введ. 28.08.2013). URL: https://gostexpert.ru/data/files/18045-2013/65454.pdf

9 Постановление Правительства Российской Федерации от 08.02.2018 № 127. URL:http://publication.pravo.gov.ru/Document/View/0001201802130006

Таблица 1.

Факторы и характеристики рисков ИБ АСУ ТП КИИ

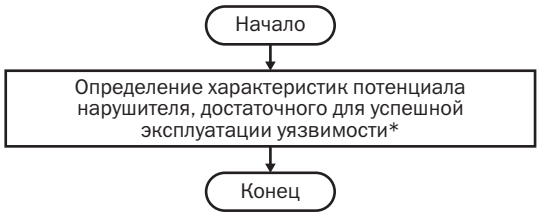
Фактор риска		Характеристики риска	Возможность количественной оценки
Величина тяжести возможных последствий от реализации опасного события		экономические последствия, вызванные нарушением критических процессов	математическое ожидание случайной величины материального ущерба
		социальные последствия, вызванные нарушением критических процессов	математическое ожидание случайной величины смертельного поражения определенного числа людей
		экологические последствия, вызванные нарушением критических процессов	математическое ожидание случайной величины аварийных выбросов в окружающую среду
		последствия угроз политической значимости объекта КИИ	—
		последствия угроз обеспечению обороны страны, безопасности государства и правопорядка	—
		степень возможного ущерба от нарушения целостности/доступности/конфиденциальности обрабатываемой в АСУ ТП информации	—
Вероятность наступления опасного события	Уязвимости системы и ее компонентов	возможная удаленность нарушителя для использования уязвимости (локальный доступ/соседняя сеть/сетевой доступ)	—
		сложность использования уязвимости	—
		требуемые привилегии для использования уязвимости	—
		необходимость взаимодействия с пользователем (необходимость действий со стороны пользователя для использования нарушителем уязвимости)	—
		возможность использования уязвимости (наличие или отсутствие кода или техники эксплуатации)	—
		уровень исправления уязвимости	—
		степень достоверности отчета о существовании уязвимости, известных технических деталей	—
	Потенциал нарушителя	время, затрачиваемое на идентификацию уязвимости и ее использование	математическое ожидание случайной величины времени обнаружения и использования уязвимости
		требующаяся техническая компетентность нарушителя для эксплуатации уязвимости	—
		знание нарушителем проекта системы и ее функционирования	—
		возможность доступа к исследуемой системе для нарушителя	—
		аппаратные средства/программное обеспечение ИТ или другое оборудование, необходимое для эксплуатации уязвимости	—

Для каждого объекта защиты АСУ ТП КИИ на основании ранее проведенной идентификации рисков ИБ АСУ ТП КИИ формируется перечень уязвимостей и проводится их анализ на основании стандарта CVSS (рис. 3).

* на основании стандарта ГОСТ Р ИСО/МЭК 18045-2013

Рис. 3. Блок-схема оценки значимости объектов защиты АСУ ТП КИИ (алгоритм 2)

Далее для каждой уязвимости оцениваются необходимые возможности нарушителя для ее успешной эксплуатации в соответствии со стандартом ГОСТ Р ИСО/МЭК 18045-2013 (рис. 4) такие как: требуемые привилегии для использования уязвимости, наличие или отсутствие кода или техники эксплуатации уязвимости, возможная удаленность нарушителя для использования уязвимости и другие.



* на основании стандарта ГОСТ Р ИСО/МЭК 18045-2013

Рис.4. Блок-схема оценки значимости объектов защиты АСУ ТП КИИ (алгоритм 3)

В результате формируется сопоставление активов, уязвимостей и возможностей нарушителей, а также определяются их характеристики, что позволяет на этапе сравнительной оценки рисков сопоставить каждой уязвимости компенсирующие меры защиты для дальнейшего снижения риска до минимального остаточного.

Согласование предлагаемого подхода к оценке рисков с практикой категорирования объектов КИИ

Согласно приказам ФСТЭК России решение о применении базовых мер обработки рисков на данный момент основывается на результатах категорирования объектов КИИ (присвоения класса защищенности АСУ). Результаты оценки рисков объектов КИИ характеризуются необходимостью согласования с результатами ранее проведенного категорирования. Например, риски объекта с третьей категорией

Таблица 2.

Сопоставление этапов управления рисками ИБ и категорирования объектов КИИ

Процесс управления рисками ИБ		Процесс категорирования объектов КИИ	
№ п/п	Этап управления рисками ИБ	№ п/п	Этап категорирования объектов КИИ
1	Установление контекста, идентификация рисков.	1	Идентификация процессов, реализующих функционирование организации.
		2	Идентификация критических процессов (которые могут привести к опасным последствиям согласно перечню показателей критериев значимости).
		3	Идентификация объектов, используемых для обработки информации, для реализации критических процессов (выделение потенциальных объектов КИИ).
		4	Формирование модели нарушителя.
		5	Формирование модели угроз (в том числе, идентификация уязвимостей).
2	Количественный и качественный анализ рисков.	6	Оценка возможных последствий в случае реализации угроз ИБ (согласно перечню показателей критериев значимости).
3	Обработка рисков.	7	Присвоение объекту КИИ определенной категории значимости или принятие решения об отсутствии такой необходимости.
4	Принятие рисков.		
5	Мониторинг и переоценка рисков.	8	Пересмотр в случае изменений значений показателей критериев значимости. Плановый пересмотр не реже, чем раз в 5 лет.

DOI: 10.21681/2311-3456-2024-1-116-123

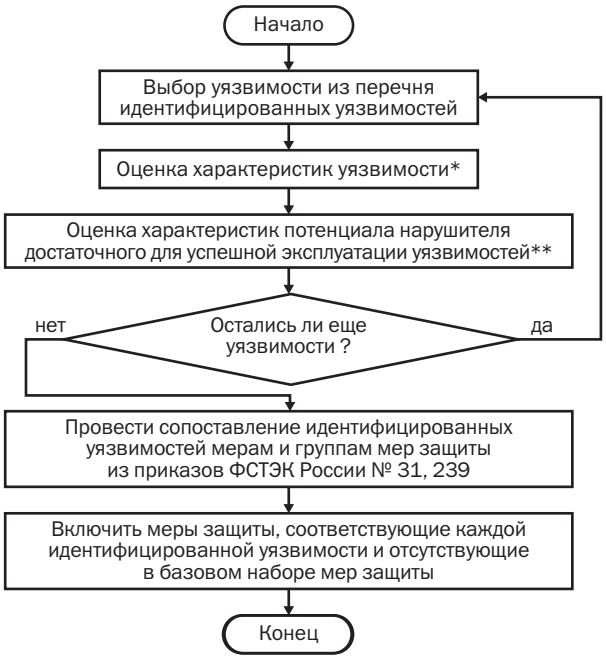
значимости не должны быть более критичными, чем риски объекта КИИ первой категории значимости. Это влечет за собой необходимость сопоставления правил категорирования с международным стандартом управления рисками ИБ. На основе анализа, проведенного в [8], ниже представлено сопоставление этапов проведения категорирования объектов КИИ согласно Постановлению Правительства № 127 и управления рисками в соответствии с ГОСТ Р ИСО/МЭК 27005-2010[10] (и проектом 2022 года[11]) (табл. 2).

Правилами категорирования определяются лишь сроки пересмотра в случае изменения значений показателей критериев значимости, которые не всегда охватывают все происходящие с системой изменения. Переоценку рисков ИБ необходимо проводить в случае модернизации и автоматизации новых процессов, применения искусственного интеллекта или биометрических технологий. Данные изменения могут не повлиять на значения показателей критериев значимости АСУ ТП КИИ, из-за чего категория значимости, а вместе с ней и применяемый базовый набор мер защиты изменены не будут, что приведет к неактуальности рисков (модели нарушителя и угроз ИБ) и нанесению ущерба субъекту КИИ.

Предложенную методику количественной и качественной оценки рисков ИБ можно включить в процесс категорирования АСУ ТП КИИ. Ниже приведена блок-схема процесса оценки рисков ИБ АСУ ТП КИИ в дополнение к проведенному категорированию объектов КИИ (рис. 5).

Идентифицированные и сопоставленные с угрозами ИБ в процессе категорирования уязвимости оцениваются по метрикам CVSS. Дополнительно оценивается потенциал нарушителя, достаточного для эксплуатации каждой из уязвимостей. Каждой угрозе сопоставляются используемые уязвимости, а каждой уязвимости – минимизирующие их меры защиты. Согласно предложенному методу оценки рисков ИБ АСУ ТП КИИ в набор мер защиты включаются те меры защиты из требований приказов ФСТЭК России № 239 и 31, что соответствуют выявленным уязвимостям и не включены в базовый набор мер защит.

Предложенная методика оценки рисков ИБ АСУ ТП КИИ может проводиться самостоятельно, так и быть включенной в дополнение к категорированию объектов КИИ. Возможность включения оценки рисков ИБ в дополнение к проводимому категорированию объектов КИИ с целью адаптации базового набора мер защиты позволяет оптимизировать и усовершенствовать обеспечение ИБ АСУ ТП КИИ.



* на основании стандарта CVSS.
** на основании ГОСТ Р ИСО/МЭК 18045-2013

Рис. 5. Оценка рисков ИБ АСУ ТП КИИ в дополнение к проводимому категорированию объектов КИИ

В качестве факторов и характеристик риска используются известные и применяемые метрики и характеристики, что не потребует дополнительного анализа, если аналогичный был уже ранее проведен.

Выводы

В рамках настоящего исследования была сформирована методика оценки рисков ИБ АСУ ТП КИИ. На примере известных инцидентов, направленных на нарушение ИБ АСУ ТП, а также на основании требований приказов ФСТЭК России была обоснована актуальность создания соответствующей методики. По итогам исследования существующих подходов количественной и качественной оценки рисков ИБ были определены достоинства и недостатки каждого из подходов применительно для АСУ ТП КИИ. Смешанный (гибридный) подход определен как компромиссный между ними. С использованием национальных и международных нормативно-методических документов и практики обеспечения информационной безопасности были определены факторы и характеристики рисков и сформированы предложения к методике количественной и качественной оценки рисков ИБ АСУ ТП КИИ. Также было проведено сопоставление процессов категорирования объектов КИИ и управления рисками ИБ и сформированы предложения по внедрению разработанной методики в дополнение к проводимому категорированию объектов КИИ.

Предлагаемый в настоящей работе подход к количественной и качественной оценке рисков ИБ АСУ

10 ГОСТ Р ИСО/МЭК 27005–2010. (введен 12.01.2011). URL: https://docs.cntd.ru/document/1200084141?ysclid=lpr75gcskw446302670

11 ГОСТ Р ИСО/МЭК 27005 (проект, первая редакция). URL: https://fstec.ru/tk-362/standarty/proekty/proekt-natsionalnogo-standarta-gost-r-iso-mek-27005

ТП КИИ ориентирован на объекты защиты и имеющиеся в системе уязвимости. Определение величины тяжести последствий основано на категории значимости объекта защиты и классе защищенности АСУ, а величина вероятности наступления опасного события определяется путем оценки уязвимостей и потенциала нарушителя. В свою очередь, потенциал нарушителя оценивается по характеристикам из ГОСТ Р ИСО/МЭК 18045-2013, предлагающего методику определения потенциала нападения нарушителя, ориентированную на имеющиеся в системе уязвимости.

Результаты проведенного анализа и выработанные рекомендации по адаптации базового набора мер защиты могут быть использованы для повышения защищенности АСУ ТП КИИ. Предложенная методика оценки рисков согласована с существующей практикой обеспечения ИБ АСУ ТП, методами управления рисками ИБ и требованиями приказов ФСТЭК России. Оценка рисков, ориентированная на объекты защиты АСУ ТП КИИ и их уязвимости, позволяет реализовать детальную оценку рисков в условиях неопределенности видов возможных нарушителей и их мотивов.

Литература

1. *Durakovskiy A. P., Gavdan G. P., Korsakov I. A., Melnikov D. A. About the cybersecurity of automated process control systems // Procedia Computer Science. 2021. № 190. P. 217–225. DOI: 10.1016/j.procs.2021.06.027.*
2. *Бабенко А. А., Магомедов Д. А. Оценка риска информационной безопасности автоматизированной системы управления технологическим процессом. Международная научно-техническая конференция «Перспективные информационные технологии» (Самара, Российская Федерация, 24–27 мая, 2021 г.). ПИТ 2021. С. 140–145.*
3. *Sembiring Z. Stuxnet Threat Analysis in SCADA (Supervisory Control and Data Acquisition) and PLC (Programmable Logic Controller) Systems // Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITTE). 2020. № 1 (2). Pp. 96–103. DOI: 10.30596/jcositte.v1i1.5116.*
4. *Geiger M., Bauer J., Masuch M., Franke J. An Analysis of Black Energy 3, Crashoverride, and Trisis, Three Malware Approaches Targeting Operational Technology Systems. 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (8–11 Sept., 2020). ETFA'2020. Pp. 1537–1543. DOI: 10.1109/ETFA46521.2020.*
5. *Maynard P., McLaughlin R., Sezer S. Decomposition and sequential-AND analysis of known cyber-attacks on critical infrastructure control systems // Journal of Cybersecurit. 2020. № 6 (1). 20 p. DOI: 10.1093/cybsec/tyaa020.*
6. *Aljohani T. M. Cyberattacks on Energy Infrastructures: Modern War Weapons // Preprint Arxiv.org (Cornell University Library). 2022. 10 p. DOI: 10.48550/arXiv.2208.14225.*
7. *Chernov D., Sychugov D. Problems of Information Security and Availability of Automated Process Control Systems. 2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM) (25–29 March 2019). ICIEAM'2019. 5 p. DOI: 10.1109/ICIEAM.2019.8743037.*
8. *Кидяева С. М., Шабурова А. В., Селифанов В. В. Вопросы организации менеджмента рисков значимых объектов критической информационной инфраструктуры // Интерэкспо Гео-Сибирь. 2022. № 6. С. 82–87.*
9. *Djurayev R. Kh., Jabborov Sh. Yu., Omonov I. I. Methods for assessing the information security of telecommunications networks. // Scientific progress. 2021. № 3. С. 73–77.*
10. *Crotty J., Daniel E. Cyber threat: its origins and consequence and the use of qualitative and quantitative methods in cyber risk assessment // Applied Computing and Informatics. 2022. 12 p. DOI: 10.1108/ACI-07-2022-0178.*
11. *Rahmani J. The main approaches to evaluating the effectiveness of applying the risk analysis and management methodology at energy company // T-Comm. 2022. № 9. Pp. 46–55. DOI: 10.36724/2072-8735-2022-16-9-46-55.*
12. *Минаков А. В. Оценка модели рисков информационной безопасности: характеристика, проблемы и перспективы // Экономика и бизнес: теория и практика. 2023. № 10–2 (104). С. 63–69. DOI: 10.24412/2411-0450-2023-10-2-63-69.*
13. *Canbolat S., Elbez G., Hagenmeyer V. A new hybrid risk assessment process for cyber security design of smart grids using fuzzy analytic hierarchy processes // Automatisierungstechnik. 2023. № 71 (9). Pp. 779-788. DOI: 10.1515/auto-2023-0089.*
14. *Харченко А. Ю., Харченко Ю. А. Анализ и определение рисков информационной безопасности // Вестник науки и образования. 2020. № 6–1 (84). С. 18–21.*
15. *Иваненко В. Г., Иванова Н. Д. Методика анализа стойкости автоматизированных систем управления технологическим процессом энергоблока АЭС к воздействию компьютерных атак // Безопасность информационных технологий. 2021. № 28 (4). С. 52–62. DOI: 10.26583/bit.2021.4.04.*

ОСОБЕННОСТИ ИДЕНТИФИКАЦИИ РАДИОЛОКАЦИОННЫХ ЦЕЛЕЙ ПРИ ОБЕСПЕЧЕНИИ БЕЗОПАСНОСТИ КРИТИЧЕСКОЙ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ

Гончаренко Ю. Ю.[1]

DOI: 10.21681/2311-3456-2024-1-124-131

Цель исследования: *систематизация особенностей процесса идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры.*

Метод исследования: *рассмотрены классические задачи оптимальной линейной фильтрации по критерию минимума среднего квадрата ошибки и по критерию максимума отношения сигнал-шум. Условия оптимальной фильтрации характеризуются спектральной плотностью полезного сигнала на входе радиоприёмного устройства.*

Результат исследования: *показано, что работное время автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры зависит от значений скорости обзора пространства, а так же от коэффициентов распознавания радиолокационных станций, необходимых для обнаружения и распознавания людей, животных, пилотируемых и беспилотных малоразмерных летательных аппаратов и других опасных целей на подходах к охраняемым объектам критической информационной инфраструктуры. Данные коэффициенты определяются с использованием функций поглощения электромагнитного излучения тканями биообъекта по определенному числу линейных интегралов, зависящих от размеров объекта и длины волн излучения.*

Научная новизна: *сформулированная задача автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры технически сводится к задаче оптимальной фильтрации радиоприёмным устройством всех отражённых сигналов.*

Ключевые слова: *радиолокационная станция, идентификация опасных целей, процесс автоматизации, физическая защита, акустический контроль, оптимальная фильтрация сигналов.*

FEATURES OF IDENTIFICATION OF RADAR TARGETS WHILE ENSURING THE SECURITY OF CRITICAL INFORMATION INFRASTRUCTURE

Goncharenko Yu. Yu.[2]

The purpose of the study: *systematization of the features of the process of identification of radar targets while ensuring the security of critical information infrastructure.*

Research method: *the classical problems of optimal linear filtering are considered by the criterion of the minimum of the mean square of the error and by the criterion of the maximum of the signal-to-noise ratio. The formulated task of automated identification of radar targets while ensuring the safety of critical information infrastructure is technically reduced to the task of optimal filtering by a radio receiver of all reflected signals. Optimal filtering conditions are characterized by the spectral density of the useful signal at the input of the radio receiver.*

Results obtained: *it is shown that the operating time of automated identification of radar targets while ensuring the safety of critical information infrastructure depends on the values of the speed of viewing space, as well as on the recognition coefficients of radar stations necessary for detecting and recognizing people,*

1 Гончаренко Юлия Юрьевна, доктор технических наук, доцент, ФГАОУ ВО «Севастопольский государственный университет», г. Севастополь, Россия. E-mail: yygoncharenko@sevsu.ru

2 Yulia Yu. Goncharenko, Dr.Sc., Associate Professor, Sevastopol State University, St. Sevastopol, Russia, E-mail: yygoncharenko@sevsu.ru

animals, manned and unmanned small-sized aircraft and other dangerous targets on approaches to protected objects of critical information infrastructure. These coefficients are determined using the absorption functions of electromagnetic radiation by the tissues of a biological object according to a certain number of linear integrals depending on the size of the object and the wavelength of radiation.

Scientific novelty: *the formulated task of automated identification of radar targets while ensuring the safety of critical information infrastructure is technically reduced to the task of optimal filtering by a radio receiver of all reflected signals.*

Keywords: *radar station, identification of dangerous targets, automation process, physical protection, acoustic control, optimal signal filtering.*

Введение

Обеспечение безопасности критической информационной инфраструктуры – актуальная проблема современности, которую призваны решать службы физической защиты этих объектов [1–3]. Они оснащены оптоэлектронными и инфракрасными средствами, системами акустического и контактного контроля, предназначенными для наблюдения за периметром охраняемых объектов и на подходах к ним [4–7]. Появление новых террористических угроз в отношении критической информационной инфраструктуры потребовало оснащение служб физической защиты специальными средствами обнаружения – радиолокационными станциями, которые обеспечивают наблюдение и освещение обстановки вокруг охраняемых объектов [8–12].

Процесс освещения обстановки включает ряд составных частей, а именно: радиолокационный поиск вокруг охраняемого объекта, обнаружение радиолокационной цели, идентификация и определение степени ее опасности, принятие решения и воздействие на обнаруженную опасную цель. Для принятия решения по варианту воздействия должностные лица действуют строго по протоколу. Проблема состоит в идентификации цели и определении степени ее опасности, при этом время, которое проходит с момента обнаружения радиолокационной цели до ее идентификации, зависит от множества объективных и субъективных факторов, включая уровень подготовки операторов радиолокационных станций, их эмоциональное и физическое состояние. Решить эту проблему можно путем автоматизированной идентификации [13–15].

Цель данной работы – систематизация особенностей процесса автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры.

Для достижения поставленной цели необходимо решить следующие задачи. Во-первых, сформулировать задачу автоматизированной идентификации при радиолокационном наблюдении вокруг объектов критической информационной инфраструктуры. Во-вторых, рассмотреть классическую задачу оптимальной линейной фильтрации по критерию минимума среднего квадрата ошибки. В-третьих, рассмотреть классическую задачу оптимальной линейной фильтрации по критерию максимума отношения сигнал-шум. В-четвертых, интерпретировать классические решения применительно к задаче автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры (КИИ).

Постановка задачи автоматизированной идентификации при радиолокационном наблюдении вокруг объектов КИИ

Главная задача радиолокационного наблюдения вокруг объектов КИИ – это обнаружение опасных целей. С позиций охраны объектов КИИ к таким целям можно отнести злоумышленников (вооруженных и невооруженных людей), двигающихся в сторону охраняемого периметра, животных и птиц, деревья и кустарники, которые могут использоваться для маскировки и отвлечения внимания операторов. Это могут быть пилотируемые (автожиры, дельтапланы, парашюты) и беспилотные летательные аппараты всех классов и типов. Именно они представляют наибольшую опасность на подходах к охраняемым объектам. Все эти цели, попав в зону радиолокационного облучения, отражают электромагнитные волны, которые принимаются антенной РЛС и обрабатываются (усиливаются, детектируются, демодулируются, преобразуются, выделяются на фоне помех). С технической точки зрения совокупность всех этих процессов принято называть фильтрацией сигналов.

Другими словами, задача идентификации опасных радиолокационных целей на подходах к охраняемому объекту КИИ технически сводится к оптимальной фильтрации принятых радиоприёмным устройством – РЛС, всех отраженных сигналов, поступающих как от опасных целей, так и от целей, создающих помехи.

Рассматривая этот процесс в общем виде, сформулируем задачу автоматизированной идентификации радиолокационных целей при обеспечении безопасности КИИ (задачу оптимальной фильтрации сигналов).

Пусть колебание $x(t)$, принятое в некотором интервале времени радиоприёмным устройством, является функцией от сигнала $S[t, \rho(t)]$ и шума $n(t)$ следующего вида:

$$x(t) = f\{S[t, \rho(t)], n(t). \quad (1)$$

Сигнал $S[t, \rho(t)]$ в общем случае может зависеть не от одного, а от нескольких параметров $\rho_i(t)$.

Допустим, что сам сигнал или его параметр – случайный процесс, и априори известны некоторые статистические характеристики сигнала и шума. Предположим также, что известен вид функции $f\{S,n\}$ (способ комбинирования сигнала и шума).

Используя эти данные, необходимо определить устройство, изображённое на рис. 1, которое оптимальным образом решит, какая реализация самого сигнала $S[t, \rho(t)]$ или его параметр $\rho(t)$ содержится в колебаниях электромагнитных волн, описываемых зависимостью (1), принятых радиоприёмным устройством РЛС.

Рис. 1. Схема оптимального фильтра

По причине наличия шума $n(t)$ с одной стороны, и случайного характера сигнала $S[t, \rho(t)]$ с другой, оценка реализации сигнала $\hat{S}[t, \rho(t)]$ или реализации его параметра не будет совпадать с истинной реализацией, поэтому возникает ошибка фильтрации.

Для количественной оценки качества фильтрации могут использоваться различные критерии, но наиболее часто применяют критерий минимума среднего квадрата ошибки и критерий максимума отношения сигнал-шум. В зависимости от принятых допущений о характере сигнала и шума сформулированная задача решается методами линейной или нелинейной фильтрации.

Если сигнал и шум взаимодействуют аддитивно, то суммарный эффект действия равен сумме входящих эффектов. Получаем, что:

$$x(t) = f\{S[t, \rho(t)] + n(t). \quad (1)$$

В этом случае при решении задачи идентификации опасных радиолокационных целей можно ограничиться линейными методами фильтрации.

Таким образом, задача автоматизированной идентификации радиолокационных целей при обеспечении безопасности КИИ, с технической точки зрения, сводится к задаче оптимальной фильтрации сигналов. Она состоит в выявлении опасных радиолокационных целей, которыми являются люди и животные, пилотируемые и беспилотные малоразмерные летательные аппараты, двигающиеся в сторону охраняемого объекта, которая решается классическими линейными методами фильтрации.

Классическая задача оптимальной линейной фильтрация по критерию минимума среднего квадрата ошибки

Пусть сигнал $S[t, p(t)] = S(t)$ и шум $n(t)$, определяющие колебания электромагнитных волн на входе приёмного устройства, описываются уравнением (2) и являются стационарными нормальными случайными процессами с известными ковариационными функциями, то есть:

$$\left.\begin{aligned} K_s(\tau) &= M\{S(t), S(t+\tau)\} \\ K_n(\tau) &= M\{n(t), n(t+\tau)\} \\ K_{sn}(\tau) &= M\{S(t), n(t+\tau)\} \end{aligned}\right\}. \quad (3)$$

Требуется определить систему, которая из принимаемого множества $x(t) = S(t) + n(t)$ с минимальной среднеквадратической ошибкой E^2 выделяет не параметр $\rho(t)$, а сам полезный сигнал $S(t)$. Иначе говоря, искомая оптимальная система должна минимизировать величину:

$$E^2 = M\left\{[\hat{S}(t) - S(t+\Delta)]^2\right\}, \quad (4)$$

где Δ – приращение времени, введенное для общности описания.

При $\Delta > 0$ оценка $\hat{S}(t)$ на входе системы должна представить (прогнозировать) значение входного сигнала $S(t)$ на время Δ.

При $\Delta = 0$ сформулированная задача сводится к выделению сигнала $S(t)$ из колебаний $x(t)$.

Строгое математическое решение сформулированной задачи для случая полубесконечного интервала наблюдения $(-\infty;t)$ было дано А. Н. Колмогоровым[3] и Н. Винером[4]. В решении было показано, что оптимальное по критерию минимума среднего квадрата ошибки устройство относится к классу линейных фильтров с постоянными параметрами.

Используя основные результаты Колмогорова-Винера, предположим, что на входе реализуемой линейной системы, представленной на рис. 2, с импульсной характеристикой $h(t)$, определяемой выражением:

$$|h(t) = \begin{cases} h(t), \text{ если } t \geq 0, \\ 0, \text{ если } t < 0, \end{cases} \quad (5)$$

воздействует стационарный случайный процесс, описываемый соотношением:

$$y(t) = \hat{S}(t) = \int_0^\infty h(\tau)x(t-\tau)d\tau\,. \quad (6)$$

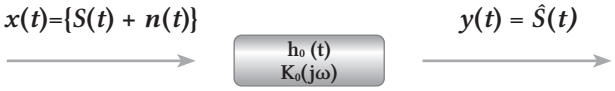


Рис. 2. Схема линейного фильтра

3 Колмогоров А. Н. Интерполирование и экстраполирование стационарных случайных последовательностей / А. Н. Колмогоров // Извещения Академии Наук, серия математическая. – М.: Академия Наук, 1941. – № 5. – С. 3–14.

4 Wiener N. Extrapolation, interpolation and soothing of stationary time series. – New-York: John Wiley, 1949. 162p.

Подставляя (6) в (4), получаем средний квадрат ошибки фильтрации:

$$E^2 = M\left\{\left[\int_0^\infty h(\tau)x(t-\tau)d\tau - S(t+\Delta)\right]^2\right\} \quad (7)$$

Выполнив в (7) ряд преобразований, получим:

$$E^2 = K_s(0) - 2\int_0^\infty h(\tau)K_{sx}(t-\tau)d\tau + \int_0^\infty\int_0^\infty h(\tau_1)h(\tau_2)K_x(\tau_2-\tau_1)d\tau_2 d\tau_1 \quad (8)$$

В выражении (8) для среднего квадрата ошибки фильтрации появляются новые составляющие. Это взаимная ковариационная функция процессов $S(t)$ и $x(t)$:

$$K_{sx}(\tau) = M\{S(t), x(t+\tau)\} \quad (9)$$

и ковариационная функция случайного процесса $x(t)$:

$$K_x(\tau) = M\{x(t), x(t+\tau)\}. \quad (10)$$

Чтобы определить импульсную характеристику $h_0(t)$ оптимального фильтра, минимизирующего средний квадрат ошибки, определяемый выражением (8), необходимо использовать один из методов вариационного исчисления.

Пусть

$$h(t) = h_0(t) + \mu g(t), \quad (11)$$

где μ – параметр, не зависящий от t; $g(t)$ – произвольная функция.

В этом случае условие минимума среднего квадрата ошибки запишется в виде:

$$\left.\frac{dE^2}{d\mu}\right|_{\mu=0} = 0. \quad (12)$$

В результате подстановки (11) в (8) условие (12) примет вид:

$$\int_0^\infty dr\left[\int_0^\infty h_0(v)dv - K_{sx}(\tau+\Delta)d\tau\right] = 0. \quad (13)$$

Соотношение (13) должно выполняться при произвольной функции $g(t)$, тогда импульсная характеристика $h_0(t)$ оптимального фильтра должна удовлетворять интегральному уравнению Фредгольма первого рода, то есть:

$$\int_0^\infty h_0(v)K_x(\tau-v)dv = K_{sx}(\tau+\Delta),\ \tau \ge 0 \quad (14)$$

Выражение (14) принято называть уравнением Винера-Хопфа. Оно является одним из основных уравнений теории линейной фильтрации.

Получили, что задача нахождения оптимального сглаживающего фильтра (при $\Delta = 0$) или оптимального прогнозирующего фильтра (при $\Delta > 0$), которая может реализоваться технически, сводится к решению уравнения (14). Это довольно сложная задача, которая обусловлена требованиями к технической реализации оптимального фильтра.

Рассмотрим частный случай. На вход фильтра поступает случайная последовательность $x(t)$, которая имеет дробно-рациональную спектральную плотность $S(\omega)$, что возможно, как правило, в результате высокочастотного детектирования электромагнитного сигнала, поступающего на вход приёмного устройства РЛС. Используя (14), получим $K_0(j\omega)$ – комплексную частотную характеристику оптимального фильтра, минимизирующего средний квадрат ошибки:

$$K_0(j\omega) = \frac{1}{2\pi\cdot f(j\omega)}\int_0^\infty e^{-i\omega r}dr\int_0^\infty \frac{S_{sx}(\Omega)}{f^*(j\Omega)}e^{j\Omega(\tau+\Delta)}d\Omega, \quad (15)$$

где

$$\left.\begin{aligned} &f(j\omega)\cdot f^*(j\omega) = |f(j\omega)|^2 = S_x(\omega),\\ &S_x(\omega) = \int_{-\infty}^\infty K_x(r)\,e^{-i\omega r}dr,\\ &S_{sx} = \int_{-\infty}^\infty K_{sx}(\tau)\,e^{-i\omega\tau}d\tau \end{aligned}\right\} \quad (16)$$

Тогда минимальное значение среднего квадрата ошибки фильтрации будет определяться выражением:

$$E^2_{\min} = \frac{1}{2\pi}\int_{-\infty}^\infty [S_s(\omega) - |K_0(j\omega)|^2\cdot S_x(\omega)]d\omega, \quad (17)$$

где $S_s(\omega) = \int_{-\infty}^\infty K_s(\tau)^{-j\omega\tau}d\tau$

Для частного случая сглаживания аддитивного множества взаимно независимых стационарных случайных процессов $S(t)$ и $n(t)$, последний из которых называют белым шумом (математическое ожидание $m_n = 0$ и кор реляционная функция $R_n(\tau) = \frac{N_0}{2}\delta(\tau)$, выражение (15) упрощается и приводится к виду:

$$K_0(j\omega) = 1 - \frac{N_0}{2S_s(\omega) + N_0}. \quad (18)$$

Под N_0 принято понимать общее число систем или число восстановлений одной и той же системы. Физический смысл N_0 состоит в количестве принятых импульсных отражений от радиолокационной цели при её идентификации. Значит, для рассматриваемого частного случая средний квадрат ошибки будет вычисляться по формуле:

$$E^2_{\min} = \frac{N_0}{2}\int_{-\infty}^\infty \ln\frac{1}{\pi}\left(1 + \frac{2S_s(\omega)}{N_0}\right)d\omega. \quad (19)$$

Практическая реализация вычислений по формуле (19) оказывается довольно громоздкой, поэтому для их упрощения не будем накладывать на оптимальный фильтр требование технической реализуемости. Тогда нижний предел в выражении (14) будет равным $-\infty$, и оно примет вид:

$$\int_0^\infty h_0(v)K_x(\tau-v)dv = K_{sx}(\tau+\Delta). \quad (20)$$

Решение (20) позволяет получить следующее выражение комплексной частичной характеристики оптимального фильтра:

$$K_0(j\omega) = \frac{S_{sx}(\omega)}{S_x(\omega)}e^{j\omega\Delta}. \quad (21)$$

Для частного случая статистически независимого сигнала $S(t)$ и белого шума $n(t)$ (21) приводится к виду:

$$K_0(j\omega) = \frac{S_{sx}(\omega)}{S_x(\omega) + S_n(\omega)} e^{j\omega\Delta}. \quad (22)$$

Допуская равенство выражений (18) и (22), описывающих одну и ту же комплексную частотную характеристику оптимального фильтра, получим:

$$1 - \frac{N_0}{2S_s(\omega) + N_0} = \frac{S_s(\omega)}{S_s(\omega) + S_n(\omega)} e^{j\omega\Delta}.$$

После предобразования последнее выражение примет вид:

$$\frac{2}{2S_s(\omega) + N_0} = \frac{e^{j\omega\Delta}}{S_s(\omega) + S_n(\omega)}, \quad (23)$$

откуда следует, что:

$$S_s(\omega) = \frac{S_n e^{-j\omega\Delta} - 0{,}5N_0}{1 - e^{-j\omega\Delta}}. \quad (24)$$

В выражении (19) в общем случае слагаемое и сомножитель $e^{-j\omega\Delta}$ является убывающей функцией, поэтому можно утверждать, что спектральная плотность сигнала будет определяться N_0, то есть:

$$S_s(\omega) = f(N_0). \quad (25)$$

Таким образом, спектральная плотность полученного сигнала при оптимальной линейной фильтрации по критерию минимума среднего квадрата ошибки определяется количеством принятых импульсных отражений от радиолокационной цели при её идентификации.

Классическая задача оптимальной линейной фильтрации по критерию максимума отношения сигнал-шум

Пусть на вход линейного фильтра с комплексной частотной характеристикой $K(j\omega)$ поступает комплексное множество $x(t)$, состоящее из полезного сигнала $S(t)$, который представляет собой случайный процесс со спектральной плотностью $S_n(\omega)$, и помехи $n(t)$ (рис. 2).

Полезный сигнал $S(t)$ статистически независим от помехи $n(t)$, форма которого заранее известна, имеет амплитудный спектр $S(j\omega)$. Тогда на выходе фильтра случайный процесс $y(t)$ будет определяться результатом преобразования сигнала $S_{лф}(t)$ и преобразования помехи $n_{лф}(t)$ линейным фильтром, то есть:

$$y(t) = S_{лф}(t) + n_{лф}(t). \quad (26)$$

Составляющая полученного сигнала на выходе фильтра будет равна:

$$S_{лф}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_s(j\omega)\, K(j\omega)\, e^{j\omega t}\, d\omega. \quad (27)$$

Дисперсия помехи на выходе фильтра:

$$\sigma^2_{лф} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_n(j\omega)\, |K(j\omega)|^2\, d\omega. \quad (28)$$

Рассмотрим величину a, представляющую собой отношение мгновенного значения полезного сигнала на входе фильтра в некоторый момент времени t, равный T, к среднеквадратическому значению выходного шума, то есть:

$$a = \frac{|S_{лф}(T)|}{\sigma_{лф}}. \quad (29)$$

С учётом (27) и (28) выражение (29) примет вид:

$$a = \frac{\frac{1}{2\pi}\int_{-\infty}^{\infty} S_s(j\omega)\, K(j\omega)\, e^{j\omega t}\, d\omega}{\sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} S_n(j\omega)\, |K(j\omega)|^2\, d\omega}}. \quad (30)$$

Линейный фильтр, максимизирующий отношение a, является оптимальным фильтром по критерию максимума отношения сигнал-шум. Его комплексные характеристики определяются формулами:

$$K_0(j\omega) = \frac{S^S(j\omega)}{S_n(\omega)} e^{-j\omega T}; \quad (31)$$

$$h_0(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K_0(j\omega)\, e^{j\omega t}\, d\omega, \quad (32)$$

где $S^S(j\omega)$ – комплексно-сопряжённая с амплитудным спектром функция входного сигнала $S(t)$; $S_n(\omega)$ – спектральная плотность помехи.

Если помеха $n(t)$, входящая в случайный процесс x(t), представляет собой стационарный нормальный Гауссовский процесс (белый шум), то выражения (31) и (32) приводятся к виду:

$$K_0(j\omega) = kS^S(j\omega)\, e^{-j\omega T}, \quad (33)$$

$$h_0(t) = kS(T-t), \quad (34)$$

где k – некоторая постоянная величина, технический смысл которой может быть определён как коэффициент усиления радиоприёмного устройства, принимающего отражённые радиолокационные сигналы.

С технических позиций коэффициент усиления приёмного устройства – постоянная величина на определённом участке полосы пропускания. В пределах всей полосы пропускания приёмного устройства коэффициент усиления изменяется в достаточно широких пределах. Это учитывается амплитудно-частотными характеристиками радиоприёмного устройства, поэтому спектральная плотность полезного сигнала на выходе линейного фильтра будет определяться степенью сопряжения коэффициента усиления амплитудному спектру входного сигнала, то есть:

$$S_s(j\omega) = K^S\, S(j\omega), \quad (35)$$

где K^S – функция, показывающая степень сопряжения (соответствия) коэффициента усиления в полосе приёмного устройства амплитудному спектру входного сигнала [17, 18].

Другими словами, спектральная плотность полезного сигнала при оптимальной линейной фильтрации по критерию максимума отношения сигнал-шум определяется степенью сопряжения (соответствия) амплитудно-частотной характеристики приёмного устройства амплитудному спектру входного полезного сигнала. Сведя в одну систему уравнений выражения (34) и (35)

$$\left.\begin{aligned} S_s(j\omega) &= f(N_0) \\ S_s(j\omega) &= K^S\, S(j\omega) \end{aligned}\right\}, \qquad (36)$$

получим условия оптимальной фильтрации сигналов, принятых радиоприёмным устройством.

Таким образом, идентификация радиолокационных целей при обеспечении безопасности КИИ технически сводится к оптимальной фильтрации принятых радиоприёмным устройством РЛС всех отражённых сигналов, поступающих как от опасных целей, так и целей, создающих помехи этому приёму. Условия оптимальной фильтрации характеризуются спектральной плотностью полезного сигнала на входе радиоприёмного устройства. Она определяется количеством принятых импульсных отражений от радиолокационной цели при её идентификации по критерию минимума среднего квадрата ошибки и степенью сопряжения амплитудно-частотной характеристики приёмного устройства амплитудному спектру входного полезного сигнала по критерию максимума отношения сигнал-шум.

Интерпретация классических решений применительно к задаче автоматизированной идентификации при радиолокационном наблюдении вокруг ядерных объектов

Совокупность сведений о целях, получаемых средствами радиолокации, принято называть радиолокационной информацией. Средства получения радиолокационной информации – это радиолокационные станции (РЛС). Другими словами, РЛС являются особой подгруппой радиотехнических систем извлечения информации и относятся к обширной группе информационных радиосистем. Процесс получения радиолокационной информации состоит из следующих этапов: обнаружение целей, измерение координат и параметров движения, разрешение и идентификация целей.

Обнаружение состоит в принятии решения о наличии или отсутствии цели в каждом временном участке пространства с минимально допустимыми вероятностями ошибочных решений.

Измерение сводится к определению координат и параметров движения цели с минимально допустимыми погрешностями. При использовании сферической системы координат обычно измеряют дальность до цели, ее азимут (на суше) или пеленг (на море) и угол места. В качестве параметров движения цели, как правило, используются направление и скорость движения. Могут выводиться производные координат либо другие параметры траектории движения, например, временное изменение расстояния (ВИР), временное изменение азимута (ВИА) и другие.

Разрешение состоит в выполнении задач обнаружения и измерения параметров выбранной цели при наличии других целей, находящихся в зоне наблюдения. В соответствии с характером движения цели различают разрешение целей по дальности, угловым координатам и скорости.

Разрешающую способность по координатам характеризуют элементарным объемом, размеры которого по дальности, в азимутальной плоскости, угломестной устанавливаются так, что наличие цели в соседнем объеме не ухудшает показатели качества обнаружения и измерения параметров цели, расположенной в центре выделенного объема. Определенный таким образом элементарный объем называют разрешаемым объемом.

Распознавание заключается в установлении принадлежности обнаруженной цели к определенному виду. В одних случаях необходимо установить принадлежность «свой – чужой» с помощью запросно-ответных устройств радиолокационного распознавания. В других необходимо выделить истинную цель на фоне ложных целей, определить характер ее движения и тип. Эти действия называются идентификацией.

Скоротечность и сложность решения каждой из задач: обнаружения, измерения, разрешения и распознавания (идентификации) для любого конечного объема пространства должно решаться за ограниченное время. В связи с этим априорное знание характеристик радиолокационных целей позволяет упростить и формализировать процесс обработки радиолокационной информации.

Основными характеристиками радиолокационных целей являются: отражающая способность, определяющая способность цели переизлучать определенную часть падающей на нее электромагнитной энергии; закон распределения и спектр флуктуаций амплитуды отраженного сигнала; закон распределения и спектр флуктуаций фазового фронта отраженного сигнала; особенности траектории движения.

Идентификация опасных радиолокационных целей описывается системой уравнений (36), одно из которых определяется общим числом принятых импульсных отражений от цели N_0, другое – отношением сопряжения амплитудно-частотной характеристики приёмного устройства амплитудному спектру полученного входного сигнала K^S. С точки зрения технической реализации число N_0 будет тем больше, чем больше количество оборотов антенны РЛС, которое зависит от скорости вращения – технической

характеристики РЛС, то есть от $N_{обор}$ – числа оборотов в минуту. Для достоверной идентификации необходимо не менее трёх регистраций, то есть трёх оборотов антенны. В ряде случаев инструкциями определяется и 4, и 5 регистраций для достоверной классификации опасной цели.

Соответствие сопряжения амплитудно-частотной характеристики приемного устройства спектру входного полезного сигнала учитывает технический параметр РЛС δ – коэффициент распознавания. Иначе говоря, время идентификации опасной цели (работное время поисковой системы) зависит от скорости вращения антенны РЛС и коэффициента распознавания, то есть:

$$t_{ПР} = f(N_{обор}, \delta) . \quad (37)$$

Необходимо отметить, что присутствующая в выражениях (36) и (37) величина δ – коэффициент распознавания, характеризует обработку отраженного от цели радиолокационного сигнала в приёмном устройстве РЛС. Как было отмечено ранее, коэффициент распознавания, безразмерная величина, лежит в основе оптимального обнаружения и показывает, во сколько раз должно быть превышение уровня или интенсивности принятого полезного сигнала над уровнем помех, действующих на вход приёмного устройства.

В то же время величина и спектр электромагнитного сигнала, отраженного от биологической цели, будут определяться суммой поглощения единичных электромагнитных лучей на максимально малых отрезках (площадках) облученной поверхности биологической цели. Особенность этого поглощения падающей электромагнитной волны специфична для разных видов (классов) биологических целей и имеет множество индивидуальных признаков, которые по интенсивности и спектру поглощения позволяют не только классифицировать биологический объект как человека (мужчину, женщину, подростка, ребенка и др.), животное (собаку, овцу, корову и пр.), но и персонализировать ее. Подобная идентификация осуществляется по интенсивности и спектру принятого электромагнитного сигнала, отраженного от биологической цели. Она реализуется посредством восстановления образа облученной цели по совокупности линейных интегралов, описывающих поглощение электромагнитной волны в каждой точке (на бесконечно малом отрезке, бесконечно малой площадке) облученной поверхности биологического объекта.

С учетом вышеизложенного можно сделать заключение, что если в приемнике РЛС реализуется оптимальный прием, то значение коэффициента распознавания должно вычисляться по формуле:

$$\frac{1}{\delta} = \int_{R^1} T(E) \exp\left\{ - \int_{L} f(xE)\, dx \right\} dE . \quad (38)$$

Таким образом, работное время автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры будет зависеть от значений скорости обзора пространства и коэффициентов распознавания радиолокационных станций. Коэффициенты распознавания радиолокационных станций, необходимые для обнаружения и идентификации людей и других опасных целей на подходах к охраняемым объектам критической информационной инфраструктуры, определяются функциями поглощения электромагнитного излучения тканями биообъекта по определенному числу линейных интегралов, зависящих от размеров объекта и длины волн излучения.

Выводы

Задача автоматизированной идентификации радиолокационных целей при обеспечении безопасности КИИ, с технической точки зрения, сводится к задаче оптимальной фильтрации сигналов. Она состоит в выявлении опасных радиолокационных целей, которыми являются люди и животные, пилотируемые и беспилотные малоразмерные летательные аппараты, двигающиеся в сторону охраняемого объекта, которая решается классическими линейными методами фильтрации.

Спектральная плотность полученного сигнала при оптимальной линейной фильтрации по критерию минимума среднего квадрата ошибки определяется количеством принятых импульсных отражений от радиолокационной цели при её идентификации.

Идентификация радиолокационных целей при обеспечении безопасности КИИ технически сводится к оптимальной фильтрации принятых радиоприёмным устройством РЛС всех отражённых сигналов, поступающих как от опасных целей, так и целей, создающих помехи этому приёму. Условия оптимальной фильтрации характеризуются спектральной плотностью полезного сигнала на входе радиоприёмного устройства. Она определяется количеством принятых импульсных отражений от радиолокационной цели при её идентификации по критерию минимума среднего квадрата ошибки и степенью сопряжения амплитудно-частотной характеристики приёмного устройства амплитудному спектру входного полезного сигнала по критерию максимума отношения сигнал-шум.

Работное время автоматизированной идентификации радиолокационных целей при обеспечении безопасности критической информационной инфраструктуры будет зависеть от значений скорости обзора

пространства и коэффициентов распознавания радиолокационных станций. Коэффициенты распознавания радиолокационных станций, необходимые для обнаружения и идентификации людей и других опасных целей на подходах к охраняемым объектам критической информационной инфраструктуры, определяются функциями поглощения электромагнитного излучения тканями биообъекта по определенному числу линейных интегралов, зависящих от размеров объекта и длины волн излучения.

Литература

1. *Гончаренко Ю. Ю. Оптимизация акустического контроля на потенциально опасных объектах / Ю. Ю. Гончаренко, М. И. Ожиганова. – Симферополь: Общество с ограниченной ответственностью «Издательство Типография «Ариал», 2023. – 120 с.*
2. *Козырева А. В. Защита объектов критической информационной инфраструктуры в 2022 году / А. В. Козырева // Вопросы устойчивого развития общества. – 2022. – № 5. – С. 1215–1223.*
3. *Информационная безопасность критической информационной инфраструктуры организаций Российской Федерации / Е. В. Данилин, В. Е. Клюев, А. П. Теленьга, А. С. Черникова // Информационная безопасность – актуальная проблема современности. Совершенствование образовательных технологий подготовки специалистов в области информационной безопасности. – 2019. – № 1(10). – С. 33–38.*
4. *Здоровцов А. Г. Оценка эффективности системы охраны периметров объектов и контроля за прилегающей территорией / А. Г. Здоровцов, А. М. Пушкарев // Альманах Пермского военного института войск национальной гвардии. – 2022. – № 4(8). С. 43–45.*
5. *Кривошея Д. Г. Средства контроля и физической защиты периметра потенциально опасных объектов / Д. Г. Кривошея, В. Л. Ефименко // Пожарная и техносферная безопасность: проблемы и пути совершенствования. – 2020. – № 1(5). – С. 368–375.*
6. *Ожиганова М. И. Архитектура безопасности киберфизической системы / М. И. Ожиганова // Защита информации. Инсайд. – 2022. – № 2(104). – С. 5–9.*
7. *Корчагин, С. И. Системы защиты периметра, 2-е переиздание, под редакцией Корчагина Сергея Игоревича // Корчагин С. И., Шанаев Г. Ф., Филатов В. В., Закиров Т. Н., Леус А. В. и др. – М.: Секьюрити Фокус, 2019. – 282 с.*
8. *Рыкунов В. Охранные системы и технические средства физической защиты объектов / Рыкунов В.: Security Focus, 2022–284 с.*
9. *Защита критической инфраструктуры государства от террористического воздействия / Е. В. Азаренко, Ю. Ю. Гончаренко, М. М. Дивизинюк, М. И. Ожиганова. – Киев: Издательство НУОУ им. Ивана Черняховского, 2018. – 82 с.*
10. *Теоретические основы интеграции технических средств охраны границы (радиолокационные станции) в единую систему безопасности / В. Е. Эчин, А. С. Мартикьян, А. К. Саматов, Т. Т. Муратбеков // Научный аспект. – 2021. – Т. 1, № 3. – С. 54–69.*
11. *Гончаренко Ю. Ю. Особенности использования стационарных радиолокационных станций для предотвращения чрезвычайных ситуаций террористического характера / Ю. Ю. Гончаренко, С. Н. Девицына // Экономика. Информатика. – 2021. – Т. 48, № 2. – С. 405–412. – DOI 10.52575/2687-0932-2021-48-2-405-412.*
12. *Гончаренко Ю. Ю. Радиолокационные станции как средство обеспечения безопасности критической информационной инфраструктуры / Ю. Ю. Гончаренко, И. Н. Карцан // Сибирский аэрокосмический журнал. – 2023. – Т. 24, № 1. – С. 90–98. – DOI 10.31772/2712-8970-2023-24-1-90-98.*
13. *Ожиганова М. И. Автоматизация выбора мер по обеспечению безопасности объекта КИИ соответствующей категории значимости при составлении модели угроз / М. И. Ожиганова, А. О. Егорова, А. О. Миронова, А. А. Головин // Энергетические установки и технологии. – 2021. – Т. 7, № 2. – С. 130–135.*
14. *Лапсарь А. П. Повышение устойчивости объектов критической информационной инфраструктуры к целевым компьютерным атакам / А. П. Лапсарь, С. А. Назарян, А. И. Владимирова // Вопросы кибербезопасности. 2022. № 2(48). С. 39–51.*
15. *Кокарева Ю. В. Социальная безопасность: теоретический и прикладной аспекты: монография / Ю. В. Кокарева, М. Н.. – Чита: ЗабГУ, 2021. – 254 с.*

CYBERSECURITY ISSUES

SCIENTIFIC PEER-REVIEWED JOURNAL

2023, № 1 (59)

Cybersecurity Issues is a research periodical scientific and practical publication specializing in information security.
Published six times a year

WWW.CYBERRUS.INFO

The journal is being published from 2013
(Registration Certificate PI No. FS 77-75239).
CrossRef number (DOI): 10.21681/2311-3456

The journal is included in the Russian list of peer-reviewed academic publications of the Higher Attestation Commission (VAK), it is registered in the Russian Science Citation Index (RSCI/RINTs) on the Web of Science (WoS) platform and holds the 1st place in its cyber security rating. The journal's articles are available in full text

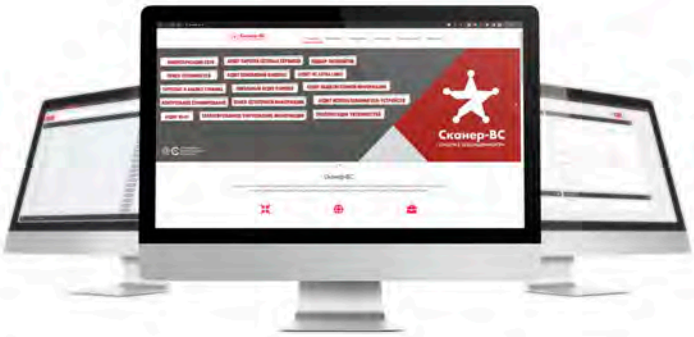
Founder and publisher
JSC «NPO «Echelon»
Postal address: Elektrozavodskaya str., 24, bld. 1, 107023, Moscow, Russia
E-mail: editor@cyberrus.info

CONTENTS

Эшелон
комплексная безопасность

Сканер-ВС
анализ защищенности

СКАНИРОВАНИЕ НА УЯЗВИМОСТИ НИКОГДА НЕ БЫЛО ТАКИМ БЫСТРЫМ!



ГК «Эшелон» представляет новый релиз системы управления уязвимостями Сканер-ВС 6. Сканер-ВС используется более чем в 5 000 организаций в России и позволяет как проводить периодическое сканирование на поиск уязвимостей, так и организовать непрерывный контроль защищенности.

Решение является ключевым компонентом, позволяющим внедрить эффективный процесс управления уязвимостями.

Скачать демо-версию «Сканер-ВС 6» (количество IP: 16, пробный период: 2 месяца) можно на сайте продукта: https://scaner-vs.ru/.

Получить техническую консультацию в группе продукта в телеграм: https://t.me/scanervs

Высокая скорость поиска

Сканер-ВС 6 обладает высокой скоростью поиска уязвимостей благодаря технологии «без скриптов»

Актуальная база уязвимостей

Ежедневно обновляемая база данных уязвимостей позволяет держать руку на пульсе последних изменений

Комплексный подход

Комплексное тестирование защищенности позволяет выявлять максимальное количество нарушений ИБ

Работа в защищенной среде

Работа в среде защищенной операционной системы Astra Linux 1.7

Отчетность

Единая среда для проведения тестирования и формирования отчетов, содержащих различную информацию в зависимости от степени детализации

Исполнение

Наличие исполнений в виде дистрибутива под Astra Linux 1.7 и LiveUSB с предустановленной ОС и с поддержкой режима сохранения изменений.

CYBERSECURITY ISSUES
VOPROSY KIBERBEZOPASNOSTI

№1 2024

DOI: 10.21681/2311-3456

- Trends and Models of Artificial Intelligence Technologies
- Network Security Techniques
- Microgrid Security Risks

www.cyberrus.info
edItor@cyberrus.Info