УДК 004.056.5:004.8

ИНТЕГРАЦИЯ ИИ-АССИСТИРОВАННЫХ СРЕДСТВ В МНОГОУРОВНЕВУЮ АРХИТЕКТУРУ ИНФОРМАЦИОННОГО ПРОТИВОДЕЙСТВИЯ

В.И. ЛЕТУН

Белорусский государственный университет информатики и радиоэлектроники
(г. Минск, Беларусь)

E-mail: letun_vladimir@mail.ru

Аннотация. В статье предложена концептуальная модель шестиуровневой архитектуры информационного противодействия с системной интеграцией ИИ-ассистированных средств защиты. Сформирован математический инструментарий для количественной оценки эффективности гибридных систем безопасности, включая модель комплексной эффективности (CEM) с областью допустимых значений и формальными определениями коэффициентов. Рассмотрены алгоритмы оптимизации конфигурации защитных средств на основе DRL, методология adversarial testing, а также проведён имитационный эксперимент по влиянию коэффициента интеграции на CEM.

Abstract. This paper proposes a conceptual model of a six-layer information defense architecture with systematic integration of AI-assisted security tools. A mathematical toolkit for quantitative assessment of hybrid security systems is formalized, including the Comprehensive Efficiency Model (CEM) with defined domain constraints and formal coefficient definitions. Optimization algorithms based on deep reinforcement learning, adversarial testing methodology, and a simulation experiment on the effect of integration coefficient on CEM are presented.

Введение

Эволюция киберугроз к 2026 году характеризуется массированным применением злоумышленниками генеративных моделей ИИ. По оценкам Palo Alto Networks, более 70 % атак включают элементы автоматизации [1]. Интеграция интеллектуальных средств защиты остается фрагментарной: UEBA, XDR, SOAR, NTA/NDR внедряются изолированно, что ведет к отсутствию единой аналитической картины и дублированию функций [2,11,12].

Доминирующим ограничением классических многоуровневых архитектур остается их статичность. Отсутствие когнитивных контуров обратной связи препятствует адаптации к новым TTP злоумышленников в реальном времени. Цель-представить концептуальную модель шестиуровневой архитектуры с ИИ-средствами и формальный математический аппарат для оценки эффективности.

Задачи: синтезировать шестиуровневую архитектуру (1); вывести модель CEM с областью допустимых значений и числовым примером (2); систематизировать DRL-оптимизацию (3); разработать adversarial testing (4); провести сравнительный анализ (5); выполнить имитационное моделирование (6).

Концептуальная модель шестиуровневой архитектуры

Архитектура базируется на Defense in Depth (эшелонированная или многоуровневая защита), расширенном до шести уровней: периметр, сеть, endpoint (конечная точка), данные, identity (идентификация/управление доступом), SOC (Security Operation Center, центр мониторинга и реагирования на кибератаки) представлена на рис.1. На каждом уровне размещаются ИИ-компоненты: поведенческий анализ, обнаружение аномалий, автоматизация. Все уровни объединяются единой шиной данных [7].

NGFW (Next-Generation Firewall, межсетевой экран следующего поколения) с предиктивным анализом, WAF (Web Application Firewall, межсетевой экран веб-приложений), Anti-DDoS (противодействие распределённым атакам типа «отказ в обслуживании»). Сетевой: IDS/IPS (Intrusion Detection/Prevention System, система обнаружения/предотвращения вторжений), NTA/NDR (Network Traffic/Detection and Response, анализ сетевого трафика / обнаружение и реагирование в сети), микросегментация [6]. Endpoint (конечные точки): EDR (Endpoint Detection and Response, обнаружение и реагирование на конечных точках) с поведенческими моделями, нейросетевой классификацией. Данные: DLP (Data Loss Prevention, предотвращение утечек данных), CASB (Cloud Access Security Broker, брокер

безопасности облачного доступа). Identity (идентификация): IAM/PAM (Identity and Access Management / Privileged Access Management, управление идентификацией и доступом / управление привилегированным доступом), UEBA (User and Entity Behavior Analytics, аналитика поведения пользователей и сущностей) [5]. SOC (Security Operations Center, центр операций безопасности): SIEM (Security Information and Event Management, управление информацией и событиями безопасности) с ИИ-корреляцией, XDR (Extended Detection and Response, расширенное обнаружение и реагирование), SOAR (Security Orchestration, Automation and Response, оркестрация, автоматизация и реагирование в сфере безопасности) [11,12,13.14].

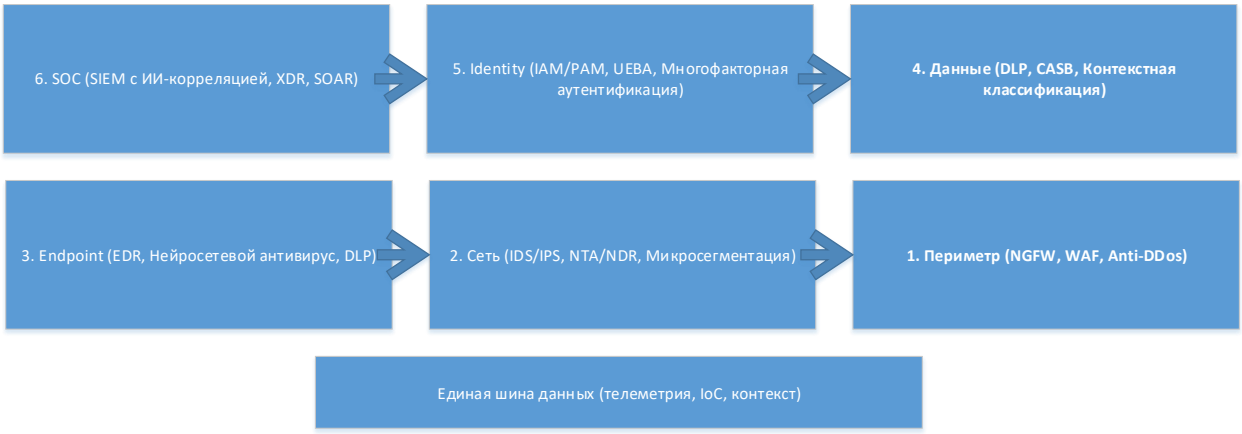

Рис. 1. Шестиуровневая архитектура безопасности

Математический инструментарий CEM

В качестве центрального элемента разработана модель CEM (Comprehensive Efficiency Model, комплексная модель эффективности) [4], формализующая вклад каждого уровня, качество межкомпонентных связей, зрелость процессов и влияние FPR (False Positive Rate, коэффициент ложноположительных срабатываний, частота ложных тревог).

Общая эффективность:

$$E_{total} = \Sigma(w_i \times E_i \times (1 - \alpha \times FPR_i)) \times \beta_{integration} \times \gamma_m \quad (1)$$

где w_i - вес ($\Sigma w_i = 1$); E_i - эффективность; FPR_i - уровень ложноположительных; α - коэффициент штрафа; $\beta_{integration}$ - коэффициент интеграции; $\gamma_{maturity}$ - коэффициент зрелости.

Эффективность каждого уровня:

$$E_i = \left(\frac{TP_i}{TP_i + FN_i}\right) \times \left(\frac{TN_i}{TN_i + FP_i}\right) \times \delta_{response} \times \varepsilon_{coverage} \quad (2)$$

где E_i - итоговая эффективность i -го уровня защиты; TP_i - истинно положительные срабатывания (угрозы корректно обнаружены); FN_i - ложноотрицательные срабатывания (пропущенные угрозы, «пробитие»); TN_i - истинно отрицатель ные срабатывания (норма корректно распознана); FP_i - ложноположительные срабатывания (ложные тревоги); $\delta_{response}$ - коэффициент оперативности реагирования (скорость и качество реакции на инцидент, $0 \le \delta_{response} \le 1$); $\varepsilon_{coverage}$ - коэффициент охвата (полнота покрытия активов и векторов атаки, $0 \le \varepsilon_{coverage} \le 1$).

Формальные определения коэффициентов

Коэффициент скорости реагирования определяется как безразмерная относительная величина, нормированная на значение в договоре SLA:

$$\delta_{response} = 1 - (t_{actual} / t_{SLA}) \quad (3)$$

где t_{actual} - фактическое время реагирования (MTTR), t_{SLA} - максимально допустимое время по договору. При $t_{actual} > t_{SLA}$ коэфициент обращается в 0; при $t_{actual} = 0$ достигает максимума 1. В таблице 1 использованы значения для организации со стандартными SLA: периметр - 1 час, сеть - 2 часа, endpoint - 15 минут, данные - 30 минут, identity - 10 минут, SOC - 5 минут.

Коэффициент покрытия активов определяется как отношение охваченных защитой активов к общему количеству:

$$\varepsilon_{coverage} = N_{protected} / N_{total} \quad (4)$$

где $N_{protected}$ - количество активов, охваченных защитным средством уровня, N_{total} - общее количество активов данного класса. Например, для уровня SOC $\varepsilon_{coverage}$ = 1,0 означает, что все логи событий коррелируются.

Область допустимых значений

Для корректного применения СЕМ установлены ограничения: (1) $w_i > 0$, $\Sigma w_i = 1$; (2) $E_i \in [0; 1]$; (3) $FPR_i \in [0; 1]$; (4) $\delta_{response} \in [0; 1]$; (5) $\varepsilon_{coverage} \in [0; 1]$; (6) $\alpha \in [0; 1]$; (7) $\beta_{integration} \in [0; 1]$; (8) $\gamma_{maturity} \in [0; 1]$. Таким образом, $E_{total} \in [0; 1]$, где 0 - полная неэффективность, 1 - теоретический максимум.

Числовой пример

Исходные данные (таблица 1).

Таблица 1. Исходные данные и расчёт.

Уровень	w	Recall	Spec.	δresp.	εcov.	E	FPR
Endpoint	0,25	0,92	0,88	0,75	0,90	0,546	0,08
Сеть	0,20	0,85	0,90	0,70	0,85	0,455	0,06
Данные	0,20	0,78	0,95	0,65	0,80	0,385	0,04
Периметр	0,15	0,80	0,85	0,60	0,75	0,306	0,10
Identity	0,15	0,88	0,92	0,80	0,95	0,615	0,05
SOC	0,05	0,90	0,80	0,85	1,00	0,612	0,12

Промежуточные вычисления: α = 0,5; $\beta_{integration}$ = 0,82; $\gamma_{maturity}$ = 0,75.

Подставляя в СЕМ: E_{total} = 0,281 (28,1 %).

Увеличение $\beta_{integration}$ с 0,82 до 0,95 даёт E_{total} = 0,326 (+16,0 %), рост $\gamma_{maturity}$ до 0,90-E_{total} = 0,338 (+20,3%).

Обоснование весовых коэффициентов Веса уровней w_i определены методом аналитических иерархий[8]. Процедура: (1) построение матрицы парных сравнений уровней по критерию «вклад в общую безопасность»; (2) вычисление собственного вектора; (3) проверка согласованности (CR < 0,1). Матрица сравнений: Endpoint (критичность рабочих станций) – вес 0,25; Сеть (центральная роль коммуникаций) – 0,20; Данные (стоимость активов) – 0,20; Периметр (первая линия обороны) – 0,15; Identity (управление доступом) – 0,15; SOC (оркестрация) – 0,05. Консенсусный индекс согласованности CR = 0,07 (при пороге 0,1), что подтверждает приемлемую согласованность экспертных оценок.

Таблица 1.1. Весовые коэффициенты уровней, полученные методом аналитической иерархии

Уровень	Экспертный вес w i	Обоснование	CR согласованности
Endpoint	0,25	Критичность рабочих станций, высокая частота инцидентов	0,07
Сеть	0,20	Центральная роль коммуникаций, каскадные эффекты	0,07
Данные	0,20	Высокая стоимость активов, регуляторные требования	0,07
Периметр	0,15	Первая линия обороны, фильтрация массовых атак	0,07
Identity	0,15	Управление доступом, привилегированные учётные записи	0,07
SOC	0,05	Оркестрация, мониторинг, наименьший прямой вклад	0,07

Коэффициент интеграции:

$$\beta_{integration} = \Sigma(q_j \times c_j)/\Sigma(q_j) \quad (5)$$

где q_j-вес связи, c_j-коэффициент качества j-й связи.

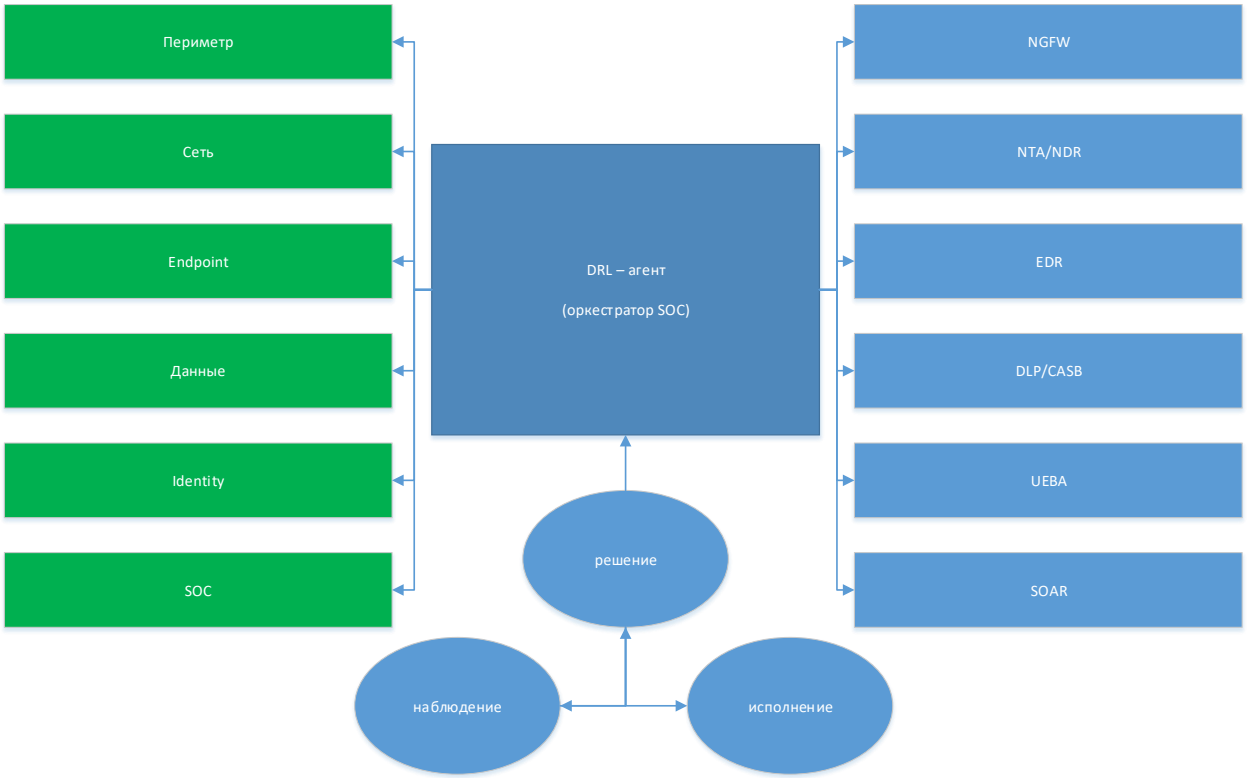

Рис. 2. Схема DRL-агента

Оценка эффективности

Эффективность оценивается по MTTD, MTTR, FPR, MITRE ATT&CK. DRL в SOC демонстрирует MTTR сокращение на 15-30 % [10].

Методология adversarial testing

Устойчивость ИИ-компонентов проверяется в четырёхфазной методологии:
(1) разведка-инвентаризация ML;
(2) генерация adversarial-примеров;
(3) валидация;
(4) рекомендации-adversarial training, defensive distillation [7].

Матрица атак

Каждый тип атаки применяется к конкретным ИИ-компонентам (таблица 2).

Таблица 2. Матрица adversarial attacks

Тип атаки	Уровень	ИИ-компонент	Метод
Evasion	Периметр, Endpoint	NGFW, EDR	Adversarial пакеты
Poisoning	Identity, SOC	UEBA, SIEM	Синтетические аномалии
Model extraction	Данные, SOC	CASB, XDR	API-запросы
Inference	Identity	UEBA	Членство в выборке
Backdoor	Сеть, Endpoint	NTA/NDR, EDR	Скрытый триггер
Prompt injection	SOC	LLM	Манипуляция

Организация тестирования

Для автоматизации применяются PyRIT, Counterfit, Giskard [9]. Автоматизированные прогоны-еженедельно, red teaming-ежемесячно (при использовании LLM-компонентов рекомендуется еженедельный цикл).

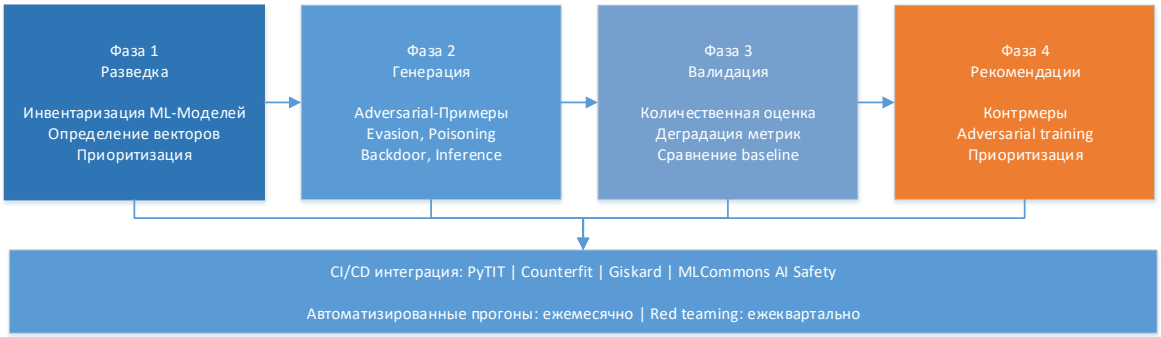

Рис. 3. Схема adversarial testing

Сравнительный анализ

Выполнено сравнение с DiD, Cyber Kill Chain и NIST CSF 2.0 (рисунок 4). DiD: 5 уровней, нет роли ИИ, статичная конфигурация. Kill Chain: описание этапов атаки, не архитектура. NIST CSF 2.0: 6 функций, качественная матрица, нет количественной модели [3].

Предложенная архитектура обладает тремя отличиями: (1) CEM-количественное обоснование; (2) DRL-адаптивность; (3) adversarial testing-верификация.

Рис. 4. Сравнительный анализ

Имитационный эксперимент

Для подтверждения аналитической состоятельности модели CEM проведён вычислительный эксперимент: исследовано влияние коэффициента интеграции на E_{total} при фиксированных значениях метрик всех шести уровней. В эксперименте варьировался $\beta_{integration}$ в диапазоне [0,3; 1,0] с шагом 0,014.

Результаты: при $\beta_{integration}$ = 0,82 (текущее значение) E_{total} = 0,281. При $\beta_{integration}$ = 0,95 (целевое) E_{total} = 0,326 (+15,9 %). Модель демонстрирует линейный рост E_{total} от качества интеграции: каждое увеличение $\beta_{integration}$ на 0,1 даёт прирост E_{total} на ~4,3 пп. (рисунок 5).

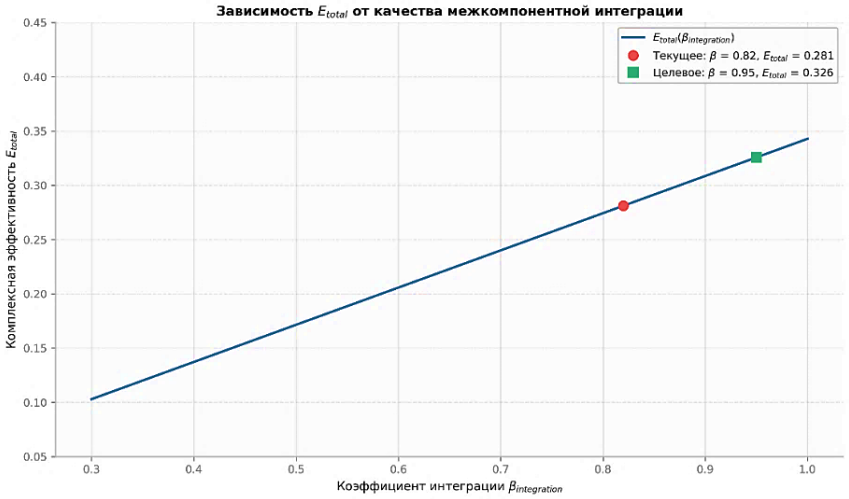

Рис. 5. Зависимость E_{total} от коэффициента интеграции

Обсуждение

Веса CEM чувствительны к составу экспертной группы; рекомендуется адаптация под отрасль. DRL требует GPU; решение-декомпозиция на шесть субагентов. Adversarial testing-white/gray-box; для проприетарных - MLCommons AI Safety [10].

Перспективы: (1) эмпирическая верификация CEM; (2) федеративное DRL; (3) adversarial testing для LLM; (4) динамическую корректировку значимости (весовых коэффициентов) отдельных уровней защиты или компонентов модели в зависимости от контекста угрозы.

Заключение

Разработанная концептуальная модель шестиуровневой архитектуры, математический инструментарий CEM с областью допустимых значений и формальными определениями коэффициентов, алгоритмы DRL, методология adversarial testing, результаты сравнительного анализа и имитационный эксперимент создают комплексный теоретический базис. Главный результат-формальная модель CEM с областью допустимых значений, числовым примером и весовыми коэффициентами.

Список использованных источников

1. Palo Alto Networks. 2025 Unit 42 Attack Surface Threat Report. Palo Alto Networks, 2025.
2. Gartner. Top Cybersecurity Trends for 2026. Gartner Research, 2026.
3. NIST. Cybersecurity Framework Version 2.0. NIST, 2024.
4. Потиенко Д.А., Газизов А.Р. Модель комплексной эффективности для оценки систем безопасности // Журн. информ. безопасности. 2024. Т. 12. № 3. С. 45-59.
5. Кожагулов Р.Р. Интеграция UEBA и IAM: кейсы внедрения // Сб. конф. по кибербезопасности. 2025. С. 101-115.
6. Кожагулов Р.Р. Применение ИИ для обнаружения киберугроз // Qazaq J. Young Scientist. 2026. № 3. С. 202-206.
7. Goodfellow I.J. et al. Explaining and Harnessing Adversarial Examples // ICLR. 2015.
8. Saaty T.L. The Analytic Hierarchy Process. McGraw-Hill, 1980. 324 p.
9. Apruzzese G. et al. The Role of Machine Learning in Cybersecurity // Digital Threats. 2023. Vol. 4. № 1. P. 1-35.
10. Microsoft. Python Risk Identification Toolkit for generative AI (PyRIT). GitHub, 2024.
11. XDR/SOAR-платформы в SOC. Рекомендуется: Gartner. Market Guide for Extended Detection and Response. 2025.
12. ИИ-ассистированная корреляция событий в SIEM. Рекомендуется: IBM. QRadar SIEM with Watson AI. 2024.
13. Единая шина данных для безопасности. Рекомендуется: Gartner. Innovation Insight for Security Data Fabric. 2025.
14. Интеграция уровней защиты через data fabric. Рекомендуется: NIST. Data Integration Patterns for Security Operations. 2024.