http://dx.doi.org/10.35596/1729-7648-2026-32-1-33-44

УДК 004.056.5:004.891:658.15

МЕТОДИКА ОЦЕНКИ ФИНАНСОВЫХ РИСКОВ ОРГАНИЗАЦИЙ НА ОСНОВЕ ВНЕДРЕНИЯ ISOLATED MULTIAGENT ARBITRATION

Е. С. ПИСКУН, А. А. АЗИЗОВ, Е. В. КРЯЧЕВ

Белорусский государственный университет информатики и радиоэлектроники (Минск, Республика Беларусь)

Аннотация. Рассмотрена проблема обеспечения снижения финансовых рисков хозяйствующих субъектов в условиях масштабного внедрения автономных интеллектуальных агентов. Показано, что существующие угрозы безопасности для систем больших языковых моделей, такие как стеганографические инъекции и поисково-дополненная генерация, трансформируются из технических инцидентов в существенные факторы операционного риска, способные нанести прямой экономический ущерб, исчисляемый миллионами долларов. Предложена методика оценки финансовых рисков на основе целевой функции полной стоимости владения, включающей операционные затраты и ожидаемые годовые потери, а также дисконтированного анализа для инвестиционного обоснования мероприятий защиты. В качестве практической реализации рассматривается архитектура Isolated Multiagent Arbitration, реализующая принцип эшелонированной защиты и изоляции генерации от исполнения и включающая модуль глубокой инспекции файлов, кастомную модель-аудитор для постгенерационного анализа ответов и механизм динамической оценки доверия к источникам в поисково-дополненной генерации.

Ключевые слова: большие языковые модели, автономные интеллектуальные агенты, промпт-инъекции, кибербезопасность, искусственный интеллект, финансовый риск, оценка стоимости, экономический эффект.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Пискун, Е. С. Методика оценки финансовых рисков организаций на основе внедрения Isolated Multiagent Arbitration / Е. С. Пискун, А. А. Азизов, Е. В. Крячев // Цифровая трансформация. 2026. Т. 32, № 1. С. 33–44. http://dx.doi.org/10.35596/1729-7648-2026-32-1-33-44.

A METHOD FOR ASSESSING THE FINANCIAL RISKS OF ORGANIZATIONS BASED ON THE IMPLEMENTATION OF ISOLATED MULTIAGENT ARBITRATION

EKATERINA PISKUN, AKBARJON AZIZOV, EGOR KRYCHEV

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Abstract. This article examines the problem of reducing the financial risks of economic entities in the context of the large-scale implementation of autonomous intelligent agents. It demonstrates that existing security threats to systems with large language models, such as steganographic injections and search-based augmentation generation, are transforming from technical incidents into significant operational risk factors capable of causing direct economic damage amounting to millions of dollars. A financial risk assessment method is proposed based on the total cost of ownership objective function, which includes operating costs and expected annual losses, as well as a discounted analysis for investment justification of security measures. The Isolated Multiagent Arbitration architecture is considered as a practical implementation. It implements the principle of layered protection and isolation of generation from execution and includes a deep file inspection module, a custom auditor model for post-generation response analysis, and a mechanism for dynamically assessing the trustworthiness of sources in search-based augmentation generation.

Keywords: large language models, autonomous intelligent agents, prompt injections, cybersecurity, artificial intelligence, financial risk, valuation, economic impact.

Conflict of interests. The authors declare that there is no conflict of interests.

For citation. Piskun E., Azizov A., Krychev E. (2026) A Method for Assessing the Financial Risks of Organizations Based on the Implementation of Isolated Multiagent Arbitration. *Digital Transformation.* 32 (1), 33–44. http://dx.doi.org/10.35596/1729-7648-2026-32-1-33-44 (in Russian).

Введение

Массовое внедрение больших языковых моделей (LLM) в критически важную бизнес-инфраструктуру с 2024 г. привело к смене парадигмы кибербезопасности: от защиты статических информационных активов к обеспечению устойчивости автономных агентных систем. Современные архитектуры наделяют агентов искусственного интеллекта (ИИ, AI-агент) правами на чтение файловых систем, выполнение API-запросов и автономное принятие решений. Это существенно расширяет поверхность атаки по сравнению с ранними реализациями генеративного ИИ, функционировавшими в изолированном режиме «текст-в-текст».

По данным McKinsey, к началу 2024 г. 65 % организаций регулярно использовали генеративный ИИ, при этом 80 % компаний увеличили инвестиции в эту область [1, 2]. В [3] прогнозируется существенный рост мировых расходов на ИИ. Параллельно фиксируется критический рост киберугроз. Согласно [4], 2023-й стал рекордным по количеству ИИ-инцидентов, включая утечки данных, дискриминационные решения и эксплуатацию уязвимостей. Совокупный ущерб от инцидентов с участием ИИ-агентов становится сопоставимым с ущербом от крупных утечек данных [4, 5], при этом скорость технологического внедрения опережает развитие процедур надзора и тестирования. Финансовые потери от подобных инцидентов повышаются на фоне роста инвестиций в ИИ-инфраструктуру. Средняя глобальная стоимость утечки данных достигла 4,88 млн долл. [5]. Уязвимость EchoLeak (CVE-2025-32711) в Microsoft 365 Copilot позволила реализовать атаку типа zero-click с эксфильтрацией данных [6, 7], единичные отравленные документы приводили к утечкам через интеграции с ChatGPT [8], атаки непрямой промпт-инъекции использовались против Slack AI и других корпоративных ассистентов [9, 10]. Автономный LLM-агент с доступом к электронной почте, RAG-хранилищам и документам в случае успешной промпт-инъекции или RAG Poisoning (Retrieval-Augmented Generation, RAG – поисково-дополненная генерация) может передавать конфиденциальные документы, раскрывать коммерческие тайны или инициировать несанкционированные действия.

Существующие защитные механизмы (RLHF и статические контент-фильтры) недостаточно эффективны против семантических обходов и обфускации [11–15]. RAG Poisoning и Jamming-атаки показывают, что даже небольшое число вредоносных документов, внедренных в базу знаний, обеспечивает высокий коэффициент успешных атак (ASR), даже если основная часть корпуса остается чистой [16–18].

Цель исследований – разработка и экономическое обоснование инвестиционного проекта по внедрению архитектуры изолированного мультиагентного арбитража (Isolated Multiagent Arbitration, IMA) для обеспечения снижения рисков финансовых потерь организации в условиях масштабного использования автономных AI-агентов.

Разработка и экспериментальная валидация эффективности архитектуры IMA

Архитектура IMA реализует подход эшелонированной защиты (Defense-in-Depth), где входящие запросы обрабатываются каскадом специализированных сервисов, взаимодействующих по REST/gRPC. Система логически разделена на три функциональных модуля, представленных на рис. 1.

Модуль 1. Deep File Inspection (DFI) – глубокая инспекция файлов. Проверка типов файлов по сигнатурам, очистка метаданных и имен, OCR-поиск стеганографии; все подозрительное блокируется по принципу Fail-Secure.

Модуль 2. Изолированный генератор (Agent Zero) дает черновой ответ, SecAudit выносит вердикт CLEAN/FLAGGED. SecAudit – трансформер-аудитор, который получает объединенный контекст (USER, RESPONSE, META, RAG) и решает, нормальный это запрос или атака, учитывая вредные инструменты и низкий trust RAG-источников. Обучающий корпус собирается из открытых наборов вредоносных/чистых промптов (AdvBench, JailbreakBench), атак на агентов (ToolEmu, AgentHarm) и сценариев RAG Poisoning/Jamming (PoisonedRAG, RobustRAG и др.), плюс корпоративные данные. Разметка идет по содержанию ответа (код, инструкции, сек-

реты), воздействию на RAG и вызовам инструментов; итоговые классы – CLEAN, JAILBREAK, EXFILTRATION, TOOL_ABUSE, RAG_POISONING, с десятками тысяч примеров на каждый.

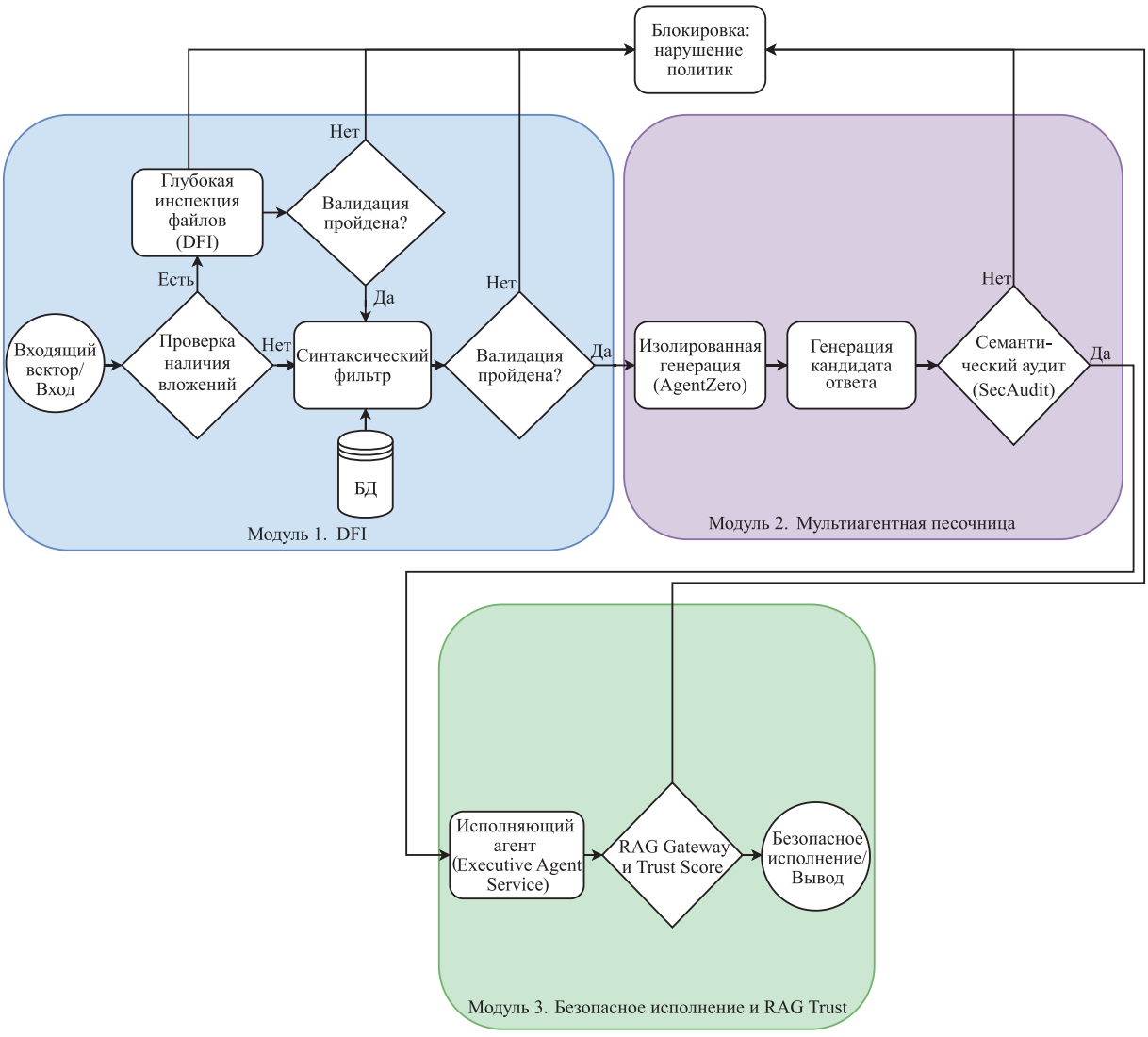

Рис. 1. Схема работы системы IMA
Fig. 1. IMA system workflow

Модель – трансформер-энкодер уровня BERT (12–24 слоя, 768–1024), дообученный как мультиклассовый/мультилейбл-классификатор с взвешенной cross-entropy и Focal loss, оптимизатор AdamW; качество контролируется по F1, особенно для EXFILTRATION и RAG_POISONING. В продакшене SecAudit работает как GPU-сервис с латентностью 300–500 мс: Agent Zero генерирует ответ, SecAudit считает распределение p(class | контекст) и по порогам p(CLEAN) и суммарного p(unsafe) выдает вердикт CLEAN или FLAGGED (с блокировкой и безопасным объяснением). Trust для RAG считается при индексации (T(d) по происхождению, свойствам домена, лингвистическим аномалиям и шаблонам poison); низко доверенные документы уходят в теневой индекс. На этапе извлечения кандидаты фильтруются по T_min, спорные помечаются RAG_TRUST=low. Возраст домена – лишь один из признаков в этой модели, а не бинарный критерий доверия.

Модуль 3. Только CLEAN-запросы попадают в Executive Agent с доступом к инструментам; контекст исполнения изолирован от проверки, все FLAGGED/блокировки логируются в SIEM, полный цикл ограничен 60 с.

В эксперименте объект защиты – автономный LLM-агент с доступом к файловой системе, внешним API и (опционально) RAG, частично недоверенному. Атакующий – подает произвольные запросы, загружает файлы, может помещать документы в RAG, но не имеет админ-доступа,

ключей и кода (black box). Цели атак: (1) jailbreak/обход выравнивания; (2) exfiltration данных из почты/файлов/RAG; (3) tool abuse для опасных действий; (4) poisoning/jamming через отравление RAG. Вне рамок – атаки на операционную систему/гипервизор/сеть, отравление предобучения, внутренние админы. Успешная атака (ASR) – вредоносный вывод LLM или принятие отравленной лжи, противоречащей базовому корпусу (PoisonedRAG).

Оценка проводилась на SecBench25 (200 сценариев, 1000 прогонов: текст, файлы, RAG). Метрики: ASR, FPR, Latency. BaselineGemini показал ASR ~16 % (160/1000), тогда как предлагаемая архитектура IMAProtection (DFI→AgentZero→SecAudit→Executive Agent + RAGtrust) продемонстрировала ASR = 0/1000; 95%-ная односторонняя верхняя граница истинного ASR ≤0,30 %. По поднаборам: текст – ≤(0,60–0,75) %, файлы – ≤(1,49–1,98) %, RAG – (≤0,99–1,19) %. Ложные блокировки не зафиксированы (FPR = 0; 95%-ная верхняя граница ≤(1,19–1,49) %). Модуль DFI заблокировал 50/50 вредоносных файлов до попадания в контекст LLM (95%-ная нижняя оценка эффективности ≥94,2 %). Средняя задержка в атакующем режиме снижена с 25,6 до 12,1 с за счет раннего отсечения атак. Ограничения: ограниченный объем выборки, отсутствие прямого сравнения с альтернативными защитами и зависимость от проприетарных LLM.

Технико-экономическая эффективность внедрения IMA

Аппарат оценки экономической эффективности внедрения IMA и снижения финансовых рисков построен от общего к частному: сначала формализуются целевая функция и алгоритм расчета, затем приводятся числовой сценарий и анализ чувствительности. При оценке экономической эффективности систем информационной безопасности рассчитывается полная стоимость владения (TCO), состоящая из прямых затрат на генерацию/исполнение ($Cost_{LLM}$), эксплуатационных затрат на защиту ($Cost_{IMA}$), ожидаемых ежегодных потерь от остаточного риска (ALE_{res}) и объема предотвращенных ежегодных потерь (ALE_{saved}). При отсутствии двойного счета (т. е. каждый элемент затрат или доходов учтен в расчетах только один раз) и раздельном учете потоков расходов совокупные издержки представляются суммой компонент (аддитивной моделью). Такой подход согласуется с технико-экономическими методиками оценки эффективности средств защиты [19] и с подходом «затраты/выгоды» в экономике информационной безопасности [20]. Для фиксации управленческих приоритетов в аддитивную модель вводятся веса w (коэффициенты значимости) слагаемых [19, 21, 22]

$$TCO = w_{LLM}Cost_{LLM} + w_{IMA}Cost_{IMA} + w_{res}ALE_{res} - ALE_{saved}. \quad (1)$$

В классическом денежном выражении сумма весов w_{LLM}, w_{IMA} и w_{res} принимается равной единице. При необходимости многокритериального выбора веса могут задаваться экспертно следующим образом:

– w_{LLM} (для $Cost_{LLM}$) позволит учесть высокую чувствительность организации к операционным издержкам; в этом случае любое снижение вычислительных ресурсов приносит значимый экономический эффект в рамках модели;

– w_{IMA} ($Cost_{IMA}$) приведет к повышению коэффициента консервативности при оценке затрат на защиту;

– w_{res} (ALE_{res}) позволит малейший остаточный риск (возможность инцидента) оценивать финансово выше его номинальной стоимости, учитывая потенциально катастрофические репутационные последствия.

Для обеспечения сопоставимости результатов и отражения управленческих приоритетов слагаемые, формирующие затратную часть TCO, взвешиваются с помощью нормированных коэффициентов w_i. Значения w_i определялись экспертным путем в соответствии с декомпозицией возможных предотвращенных ежегодных финансовых потерь организации: $w_{LLM} = 0,75$; $w_{IMA} = 0,15$ и $w_{res} = 0,10$. Отрицательное слагаемое ALE_{saved} учитывается в модели без весового коэффициента, так как представляет собой прямой объем предотвращенного ущерба, уменьшающий совокупные издержки [23, 24].

Для оценки экономической эффективности внедрения архитектуры IMA был разработан сценарий для предприятия среднего масштаба. Расчетная модель базировалась на следующих вводных параметрах:

– масштаб внедрения: 1000 активных пользователей N_{users}, генерирующих в среднем 20 запросов в сутки при 250 рабочих днях в году;

– общий объем нагрузки: $N_{year} = 1000 \cdot 20 \cdot 250 = 5 \cdot 10^6$ запросов в год.

Структура затрат разделяется на первоначальные (InitialEx) и последующие ежегодные операционные затраты (OpEx).

InitialEx (Year) включают расходы на R&D, развертывание микросервисной архитектуры и первичную интеграцию IMA-контура. Исходя из оценки трудозатрат команды (3–5 инженеров на 6–9 месяцев) и настройки инфраструктуры, CapEx оценивается в 150 000 долл. в год.

OpEx (Annual) включают:

– LLM-инференс: при стоимости сложного запроса (Agent Zero + Исполнитель) $P_{req} \approx 0{,}016$ долл., затраты на токены составляют ~80 000 долл. в год;

– инфраструктуру и фонд оплаты труда: амортизация мощностей (GPU/CPU для DFI/SecAudit) и эксплуатационные расходы (1–2 FTE MLOps/SecOps) оцениваются в 100 000 долл. в год;

– итого OpEx ≈ 180 000 долл. в год.

Алгоритм количественного анализа рисков, интегрированный в модель оценки инвестиционной эффективности внедрения IMA, имеет следующий вид [21, 24, 25]:

– определение ущерба одного значимого инцидента SLE (Single Loss Expectancy), долл.;

– оценка частоты (интенсивность) критических попыток атак на агентную систему ARO_{att} (1/год), экспериментальная оценка вероятности успеха попытки ASR*;*

– получение ожидаемого числа успешных инцидентов в год (ARO, 1/год)

$$\mathrm{ARO} = \mathrm{ARO}_{\mathrm{att}} \cdot \mathrm{ASR}; \tag{2}$$

– расчет ожидаемых годовых потерь, позволяющий перевести абстрактные угрозы в ожидаемые денежные потери, что необходимо для сопоставления с затратами на внедрение

$$\mathrm{ALE} = \mathrm{SLE} \cdot \mathrm{ARO}; \tag{3}$$

– определение предотвращенного ущерба (годовой эффект) от внедрения IMA, которая позволяет оценить разницу между потерями без защиты и с защитой, т. е. это «доходная» часть проекта

$$\Delta\mathrm{ALE} = \mathrm{ALE}_{\mathrm{base}} - \mathrm{ALE}_{\mathrm{IMA}}; \tag{4}$$

– формирование чистого денежного потока проекта на горизонте H лет

$$\mathrm{CF}_t = \Delta\mathrm{ALE} - \mathrm{OpEx}_t; \tag{5}$$

– расчет чистого дисконтированного дохода (Net Present Value, NPV)

$$\mathrm{NPV} = -\mathrm{InitialEx} + \sum_{t=1}^{5} \frac{\left(\mathrm{ALE}_{\mathrm{base}} - \mathrm{ALE}_{\mathrm{IMA}}\right) - \mathrm{OpEx}_t}{\left(1+r\right)^t}, \tag{6}$$

где InitialEx – инвестиционные затраты (Initial Investment), долл.

Экономическая эффективность предложенной архитектуры IMA базируется на уменьшении показателя ALE. Технические испытания на тестовом наборе продемонстрировали уменьшение ASR с 16,0 до 0,3 %, что соответствует потенциальному снижению ожидаемого ущерба до 98,1 %. Однако в финансовой модели используется сценарный коэффициент эффективности снижения риска k (0–1), %, отражающий реализуемую долю предотвращаемого ущерба с учетом остаточного риска, человеческого фактора и неполного охвата сценариев; в базовом сценарии $k = 0{,}5$.

На основании [26] был рассчитан индекс рентабельности (Profitability Index, PI) для ранжирования проектов

$$\mathrm{PI} = \frac{\sum_{t=1}^{5} \frac{\mathrm{CF}_t}{\left(1+r\right)^t}}{\mathrm{InitialEx}}. \tag{7}$$

Индекс доходности дисконтированных инвестиций ID, показывающий относительную отдачу именно на вложенный капитал (сверх возврата самих инвестиций) [26], определяли по формуле

$$\mathrm{ID} = \mathrm{PI} - 1. \tag{8}$$

Для анализа структуры последствий, характерных для финансового сектора, используются экспертно-аналитические данные центра InfoWatch [27]. Согласно материалам исследования,

последствия инцидентов в финансовых организациях смещены в сторону косвенного ущерба: доля прямых хищений денежных средств остается относительно низкой, в то время как основной объем рисков (более 55 %) связан с компрометацией конфиденциальной информации (персональных данных и коммерческой тайны), что влечет за собой критические репутационные издержки и долгосрочные затраты на ликвидацию последствий утечек.

В табл. 1 приведена структура предотвращенных финансовых потерь, основанная на анализе законодательства Республики Беларусь. Исходные данные для расчета принимались следующие: SLE = 4,88 млн долл. [5] (может быть откалиброван под конкретную организацию с учетом стоимости ее активов); ARO = 0,1 (значение выглядит консервативным или заниженным, так как показывает один значимый инцидент раз в 10 лет, но использование консервативного значения в сочетании с глобальным медианным показателем SLE позволяет сформировать нижнюю границу оценки экономического эффекта от внедрения архитектуры IMA); k = 50 % (базовый сценарий для финансовой модели). Сумма предотвращенной потери: 4 880 000 · 0,1 · 50 % = 244 000 долл./год. Очевидно, что внедрение IMA обеспечивает не только защиту от прямых финансовых убытков, но и снижает значительные юридические риски.

Таблица 1. Структура предотвращенных ежегодных финансовых потерь
Table 1. Structure of annual prevented financial losses

Категория риска	Описание (согласно законодательству РБ)	Доля в общем эффекте, %	Сумма, долл./год
Декомпозиция суммы предотвращенной потери			
Прямые потери	Высокая доля прямых потерь обоснована следующими факторами: технические затраты: в случае компрометации LLM-системы или RAG-базы (базы знаний) организация несет колоссальные расходы на аудит кода, очистку данных и переобучение/донастройку моделей; операционный простой: финансовый сектор критически зависит от непрерывности процессов. Стоимость часа простоя интеллектуальных фронт-офисных систем (например, скоринга или клиентской поддержки) оценивается в десятки тысяч долларов; стоимость данных: утечка интеллектуальной собственности или баз клиентов имеет прямую рыночную оценку, которая в 75 % случаев формирует основной объем материального ущерба (SLE) [5, 21]	75	183 000
Комплаенс-штрафы	Отражают регуляторную среду РБ: Закон № 99-З «О защите персональных данных»: любая успешная атака, приведшая к утечке, влечет за собой административную (а в ряде случаев и уголовную) ответственность для должностных лиц [28]; учитывая требования ст. 23.7 КоАП РБ [29] и нормы постановления НБ РБ № 351 [30], регуляторный риск (штрафные санкции и меры надзорного реагирования) принимается как консервативная величина в размере 15 % от совокупного SLE, что соответствует экспертным оценкам БелИСА [27] и мировым бенчмаркам стоимости инцидентов [5]	15	36 600
Регуляторные риски	Косвенные, но неизбежные расходы [29]: внеплановый аудит: после крупного инцидента организация обязана провести глубокую проверку ИБ-инфраструктуры с привлечением внешних сертифицированных лабораторий. Это дорогостоящая процедура, стоимость которой фиксирована; судебные издержки: сюда включены расходы на юридическое сопровождение претензий от пострадавших клиентов	10	24 400

Согласно [29] (табл. 1), штрафы за нарушение законодательства о защите персональных данных являются существенными для бизнеса, однако наибольший финансовый урон наносят сопутствующие издержки: обязательный внеплановый аудит информационной безопасности и возможные судебные иски. Расчеты показывают, что около 25 % экономического эффекта системы IMA (или ~61 тыс. ежегодно в базовом сценарии) формируется именно за счет предотвращения регуляторных и комплаенс-издержек, что делает проект критически важным для организаций, работающих с чувствительными данными граждан Беларуси. Поскольку расчеты ведутся в долларах, то внедрение IMA снижает зависимость от платных зарубежных систем безопасности (облачных WAF, AI Guardrails). Это соответствует государственной политике импортозамещения программного обеспечения [31–33].

Для учета стоимости денег во времени применялся метод дисконтированных денежных потоков (DCF) со ставками дисконтирования r = 8 %, r = 12 % и r = 16 %, соответствующими принципам риск-ориентированного планирования, изложенным в [34], и методическим рекомендациям [26]. В табл. 2 приведен анализ того, как проект IMA выглядит при различных ставках дисконтирования.

Таблица 2. Сравнительный анализ эффективности проекта IMA при различных ставках дисконтирования (H = 5 лет)
Table 2. Comparative analysis of the effectiveness of the IMA project at different discount rates (H = 5 years)

Ставка дисконтирования, %	Тип сценария	NPV, тыс. долл.	Вывод для инвестора
8	Социально-государственный	105,5	Положительный NPV и быстрая окупаемость при низкой стоимости капитала. Рекомендуется для реализации в рамках программ цифровизации госсектора и критической инфраструктуры. Проект генерирует значительный общественный эффект и снижает системные риски при низкой стоимости фондирования
12	Базовый (умеренный)	80,7	Высокая инвестиционная привлекательность. Проект обеспечивает устойчивый доход, значительно превышающий средневзвешенную стоимость капитала. Срок окупаемости и рентабельность (PI > 1) соответствуют стандартам стабильных финансовых институтов. Дисконтированный срок окупаемости – на рубеже второго-третьего годов
16	Венчурный (агрессивный)	59,6	Целесообразность подтверждена. Несмотря на высокую премию за риск, проект остается прибыльным. Рекомендуется для частных инвесторов и венчурных фондов. Риск окупаемости нивелируется высокой технической эффективностью

На рис. 2 представлена динамика накопленного DCF проекта IMA на горизонте пяти лет. Согласно методике [26], ключевыми индикаторами эффективности внедрения IMA выступают:

– дисконтированный срок окупаемости: точка пересечения кривой с осью абсцисс достигается на рубеже второго-третьего годов эксплуатации (DPP ≈ 2,9 года при r = 12 %). Короткий для наукоемких IT-проектов срок окупаемости обусловлен высокой стоимостью предотвращаемых рисков по сравнению с затратами на разработку и внедрение;

– индекс рентабельности: на конец пятого года индекс рентабельности PI ≈ 1,54 (при r = 12 %). Значение PI > 1 подтверждает целесообразность инвестиций: каждый вложенный доллар (или эквивалент в бел. руб.) генерирует около 0,54 долл. США чистой приведенной прибыли за счет снижения ожидаемых потерь от киберинцидентов.

Эффект уменьшения потерь относится к снижению непроизводительных операционных затрат и времени обслуживания под атакующей нагрузкой. В IMA вредоносные запросы отсекаются на более ранних и дешевых стадиях (DFI/AgentZero/SecAudit) и не доходят до дорогостоящего исполнения (Executive Agent и вызовы инструментов). Поэтому при росте доли атакующих запросов уменьшаются средняя задержка и потребление вычислительных ресурсов. В эксперимен-

тах для атакующих запросов средняя задержка уменьшилась на 52,7 % (с 25,6 до 12,1 с), что снижает риск отказа в обслуживании на уровне бизнес-логики. Значение ID = 0,54 свидетельствует о том, что каждый вложенный в систему IMA доллар (в приведенных ценах) приносит организации 0,54 долл. США чистой прибыли сверх возврата вложенных средств. Согласно [26], проект можно признать эффективным, так как ID > 0.

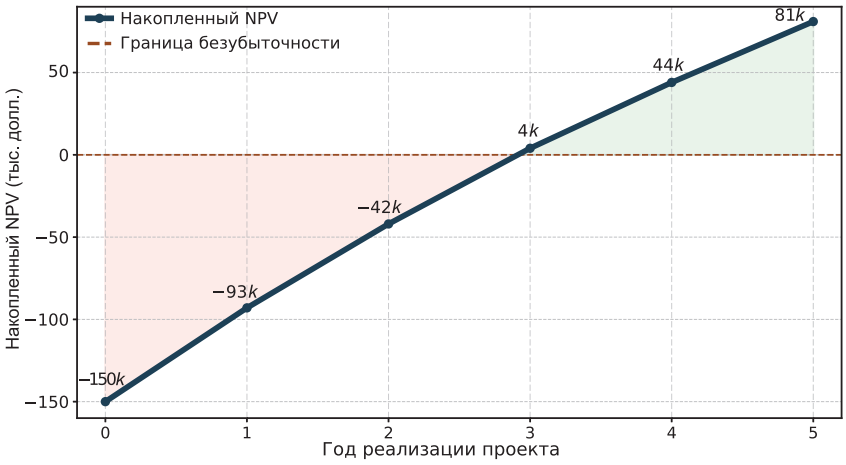

Рис. 2. Динамика чистого дисконтированного дохода
Fig. 2. Dynamics of net present value

Для проверки устойчивости результата к неопределенности входных параметров выполняли вероятностный факторный анализ как метод снижения размерности множества факторов, влияющих на TCO/NPV. В исходную модель были включены факторы 1 (показатели технической эффективности контура защиты η) и 2 (компоненты стоимости инцидента SLE), перечисленные в табл. 3 [35–37].

Таблица 3. Матрица факторных нагрузок (метод главных компонент, varimax-ротация)
Table 3. Factor loadings matrix (principal component method, varimax rotation)

№ пп	Фактор влияния	Фактор 1	Фактор 2	Общность h^2
	Техническая эффективность			
1	Точность детекции семантических атак	0,92	0,11	0,86
2	Доля ложноположительных срабатываний	0,88	0,15	0,80
3	Задержка системы арбитража	0,81	0,08	0,66
4	Коэффициент доступности агентов-цензоров	0,79	0,21	0,67
	Экономическая тяжесть инцидента			
5	Прямые убытки от компрометации данных	0,14	0,94	0,90
6	Размер регуляторных штрафов (закон № 99-З)	0,09	0,89	0,80
7	Репутационные потери (отток клиентов)	0,18	0,85	0,75
8	Стоимость восстановления RAG-инфраструктуры	0,22	0,78	0,66
	Операционная зрелость IT-инфраструктуры			
9	Расходы на вычислительные ресурсы (GPU/API)	0,45	0,38	0,35
10	Затраты на фонд оплаты труда специалистов поддержки	0,39	0,41	0,32
	Итого в сумме	**4,21**	**3,85**	
	Доля объясненной дисперсии, %	**42,1**	**38,5**	

Анализ представленной в табл. 3 матрицы факторных нагрузок позволяет сделать следующие выводы.

Во-первых, применение метода главных компонент позволило выделить два доминирующих фактора, суммарная доля объясненной дисперсии которых составила 80,6 % (42,1 % + 38,5 %). Это существенно больше общепринятого в экономических исследованиях порога (70 %) и подтверждает высокую информативность модели.

Во-вторых, выявлено четкое разделение исходных переменных на две группы:

– фактор η, аккумулирующий в себе точность детекции и надежность системы арбитража (нагрузки по переменным 1–4 – более 0,79);

– фактор SLE, объединяющий прямые убытки, репутационный ущерб и регуляторные штрафы согласно [28] (нагрузки по переменным 5–8 – более 0,78).

В-третьих, малые значения h^2 для вычислительных ресурсов и затрат на специалистов (переменные 9, 10) позволяют исключить их из дальнейшего анализа устойчивости без потери точности прогноза. Однако это не означает их исключения из общей формулы TCO, а лишь подтверждает их низкую волатильность при изменении параметров безопасности.

Суммарная доля объясненной дисперсии составила 80,6 %, что, согласно опроснику Кеттелла [38], свидетельствует о высокой репрезентативности перечисленных факторов. Оставшаяся часть дисперсии (19,4 %) относится к специфической вариативности отдельных показателей и случайным факторам, что допустимо для технико-экономических моделей управления рисками. Таким образом, факторный анализ математически обосновывает сведение многомерной задачи оценки рисков к двум ключевым осям чувствительности, что делает методику оценки финансовой устойчивости системы IMA прозрачной и пригодной для оперативного управления.

На рис. 3 с учетом вероятностной природы киберрисков представлен анализ чувствительности NPV к изменению ключевых факторов – стоимости одного инцидента SLE и коэффициента эффективности снижения риска *k*.

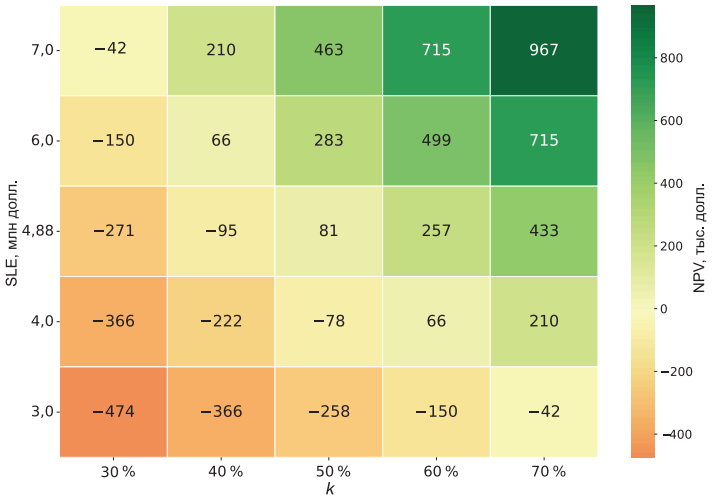

Рис. 3. Анализ чувствительности NPV
Fig. 3. NPV sensitivity analysis

Итоговые показатели экономической эффективности внедрения системы IMA приведены в табл. 4.

Таблица 4. Показатели экономической эффективности внедрения системы IMA
Table 4. Economic efficiency indicators for the implementation of the IMA system

Наименование показателя	Условное обозначение	Значение	Интерпретация согласно методике Минэкономики
Инвестиционные затраты, долл.	InitialEx	150 000	Единовременные затраты на разработку и интеграцию (год 0)
Чистый дисконтированный доход, долл.	NPV	80 706	Проект эффективен (NPV > 0) при базовых параметрах модели (r = 12 %, горизонт – 5 лет)
Индекс рентабельности	PI	1,54	На каждый вложенный 1 долл. США проект генерирует 1,54 долл. США дисконтированных выгод
Индекс доходности дисконтированных инвестиций	ID	0,54	Чистая отдача на капитал сверх возврата инвестиций составляет 54 % (ID = PI − 1)
Дисконтированный срок окупаемости	DPP	2,9 года	Соответствует наукоемким IT-проектам и позволяет уложиться в типовой жизненный цикл программного продукта в РБ

Анализ подтверждает, что в условиях цифровой трансформации Беларуси внедрение системы IMA соответствует стратегическим целям цифровизации, закрепленным в [31], обеспечивая безопасную среду для функционирования интеллектуальных агентов, а инвестиции в IMA являются «защитными активами»: они не только обеспечивают технологический суверенитет, но и гарантируют возврат инвестиций через предотвращение катастрофических убытков, превышающих стоимость разработки в десятки раз. Помимо количественных финансовых показателей, интеграция IMA в процессы корпоративной безопасности в соответствии с [39] обеспечивает снижение остаточного риска (Residual Risk) за счет трехуровневого контроля:

1) Data Layer: санитарная обработка входящих данных (модуль DFI);

2) Execution Layer: валидация логики и действий агентов (SecAudit + Executive Agent);

3) Knowledge Layer: оценка доверия к источникам в RAG-системах.

Для бизнеса это:

– обеспечение масштабируемости агентного ИИ без экспоненциального роста рисков и аудируемости решений ИИ, что критически важно для соответствия регуляторным требованиям в финансовом и государственном секторах, включая требования законодательства о персональных данных РБ;

– высокий экономический эффект, так как каждый заблокированный на уровне DFI запрос стоит доли цента (порядка 0,0001 долл.), в то время как полный цикл генерации и исполнения LLM-агентом (модели frontier-класса) может стоить около 0,016 долл. При массированных атаках (например, 100 000 вредоносных запросов в день) система не только защищает данные, но и снижает расходы на токены, предотвращая бесполезную работу дорогостоящих моделей. Экономия может составлять тысячи долларов в месяц только на вычислительных ресурсах. Данное решение соответствует приоритетам развития цифровой экономики Республики Беларусь, закрепленным в [31].

Заключение

1. Разработанная система IMA, направленная на снижение финансовых и операционных рисков, возникающих при эксплуатации автономных LLM-систем, повышает безопасность автономных LLM-агентов и одновременно дает ощутимый экономический эффект. За счет снижения успешности атак и ускоренной обработки подозрительных запросов (менее 52,7 % времени) система уменьшает ожидаемые ежегодные потери при средней стоимости инцидента SLE = 4,88 млн долл. и базовых затратах CapEx = 150 000 долл., OpEx = 180 000 долл. в год.

2. Проведенный анализ показал выход на безубыточность на рубеже второго-третьего годов и положительный дисконтированный доход к пятому году даже при умеренной эффективности снижения риска (k = 0,5). Значение индекса доходности дисконтированных инвестиций позволяет рассматривать IMA не как расход, а как инвестиционный инструмент, укрепляющий долгосрочную устойчивость бизнеса в условиях цифровой трансформации.

3. Разработанная система соответствует стратегическим приоритетам Республики Беларусь в области импортозамещения программного обеспечения и обеспечения технологического суверенитета в условиях цифровой трансформации экономики. Ее внедрение позволяет организациям не только соответствовать жестким регуляторным требованиям в области защиты персональных данных, но и рассматривать инвестиции в кибербезопасность искусственного интеллекта как возвратный актив с доказанной доходностью на капитал (ID = 0,54).

Список литературы / References

1. Singla A., Sukharevsky A., Yee L., Chui M., Hall B. (2024) *The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value*. USA, McKinsey & Company Publ. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (Accessed 24 May 2024).
2. Brier P., Thibaud A.-L., Marandon A., Shah H., Roberts Dr. M., Jones S. (2024) Harnessing the Value of Generative AI. *Capgemini Research Institute*. Available: https://www.capgemini.com/wp-content/uploads/2024/05/Final-Web-Version-Report-Gen-AI-in-Organization-Refresh.pdf (Accessed 15 August 2024).
3. *Gartner Says Worldwide AI Spending Will Total $1.5 Trillion in 2025*. Stamford, Connecticut, 2025. Available: https://www.gartner.com/en/newsroom/press-releases/2025-09-17-gartner-says-worldwide-ai-spending-will-total-1-point-5-trillion-in-2025 (Accessed 10 October 2025).
4. *2023 Was a Record Year for AI Incidents*. Surfshark Research, 2024. Available: https://surfshark.com/research/chart/ai-incidents-2023 (Accessed 12 February 2024).

5. *Cost of a Data Breach Report 2024*. IBM Security, 2024. Available: https://www.ibm.com/reports/data-breach (Accessed 20 July 2024).
6. *CVE-2025-32711 Detail.* NIST, National Vulnerability Database, 2025. Available: https://nvd.nist.gov/vuln/detail/CVE-2025-32711 (Accessed 20 May 2025).
7. *Inside CVE-2025-32711 (EchoLeak): Prompt Injection Meets AI Exfiltration*. Hack the Box, 2025. Available: https://www.hackthebox.com/blog/cve-2025-32711-echoleak (Accessed 22 May 2025).
8. Burgess M. (2025) A Single Poisoned Document Could Leak 'Secret' Data Via ChatGPT. *Wired*. Available: https://www.wired.com/story/chatgpt-poisoned-document-data-leak/ (Accessed 14 March 2024).
9. *Slack AI Can Leak Private Data Via Prompt Injection*. The Register, 2024. Available: https://www.theregister.com/2024/08/21/slack_ai_prompt_injection/ (Accessed 25 August 2024).
10. *How Microsoft Defends Against Indirect Prompt Injection Attacks*. Microsoft Security Response Center, 2025. Available: https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks (Accessed 30 July 2025).
11. Zou A., Wang Z., Kolter J. Z., Fredrikson M. (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models (GCG). *arXiv Preprint*. Available: https://arxiv.org/abs/2307.15043 (Accessed 15 January 2024).
12. Robey A., Wong E., Hassani H., Pappas G. J. (2023) SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv Preprint*. Available: https://arxiv.org/abs/2310.03684 (Accessed 20 January 2024).
13. Huang D., Shah A., Alexandre A., David W., Chawin S. (2025) Stronger Universal and Transferable Attacks by Suppressing Refusals. *NAACL*. Available: https://doi.org/10.18653/v1/2025.naacl-long.302 (Accessed 10 May 2025).
14. Su J., Kempe J., Ullrich K. (2024) Mission Impossible: A Statistical Perspective on Jailbreaking LLMs. *arXiv*. Available: https://arxiv.org/abs/2408.01420 (Accessed 1 September 2024).
15. Zeng Y., Lin H., Zhang J., Yang D., Jia R., Shi W. (2024) How Johnny Can Persuade LLMs to Jailbreak Them. *arXiv*. Available: https://arxiv.org/abs/2401.06373 (Accessed 15 February 2024).
16. Zou W., Geng R., Wang B., Jia J. (2025) PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. *Proceedings of USENIX Security*. Available: https://arxiv.org/abs/2402.07867 (Accessed 12 March 2025).
17. Xiang Ch., Wu T., Zhong Z., Wagner D., Chen D., Mittal P. (2024) Certifiably Robust RAG against Retrieval Corruption. *arXiv Preprint*. Available: https://arxiv.org/abs/2405.15556 (Accessed 10 June 2024).
18. Shafran A., Schuster R., Shmatikov V. (2024) Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents. *arXiv Preprint*. Available: https://arxiv.org/abs/2406.05870 (Accessed 15 July 2024).
19. Gaidamakin N. A. (2025) Methodology of Expert-Analytical Analysis of Technical and Economic Efficiency of the Information Security System of an Enterprise Based on Comparison with "Best Practices". *Voprosy Kiberbezopasnosti*. (5), 149–161 (in Russian).
20. Kozyr N. S. (2023) Costs and Benefits of Business Information Security. *Management*. 11 (4), 110–118 (in Russian).
21. Astakhov A. M. (2017) *The Art of Information Risk Management*. Saratov, Profobrazovanie Publ. (in Russian).
22. Kovaleva N. V. (2021) Methods of Financial Risk Assessment and Possibilities of Their Application in Modern Economic Conditions. *Consumer Cooperativ*es. 1 (72), 34–38 (in Russian).
23. Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., et al. (2008) *Global Sensitivity Analysis: The Primer*. Chichester, John Wiley & Sons, Ltd.
24. Lukasevich I. Ya. (2016) *Financial Management*. Moscow, National Education Publ. (in Russian).
25. Petrenko S. A., Simonov S. V. (2009) *Management of Information Risks. Economically Justified Security*. Moscow, DMK Press (in Russian).
26. Methodological Recommendations for Assessing the Efficiency of Investment Projects. *Approved by the Ministry of Economy, Ministry of Finance, and Ministry of Architecture and Construction, No 158/104/246. National Register of Legal Acts of the Republic of Belarus, 2005, No 158, 8/13148* (in Russian).
27. Information and Network Infrastructure Protection. *InfoWatch, 2025*. Available: www.infowatch.ru (Accessed 12 February 2026) (in Russian).
28. On Personal Data Protection. *Law of the Republic of Belarus, May 7, 2021, No 99-Z. National Register of Legal Acts of the Republic of Belarus, 2021, No 2/2819* (in Russian).
29. Code of the Republic of Belarus on Administrative Offenses, January 6, 2021, No 91-Z (Amended October 11, 2024, No 37-Z). *National Register of Legal Acts of the Republic of Belarus, 2021, No 2/2811* (in Russian).
30. On Approval of the Instruction on Requirements for Ensuring Information Security in the Banking System of the Republic of Belarus. *Resolution of the Board of the National Bank of the Republic of Belarus, November 25, 2021, No 351. National Register of Legal Acts of the Republic of Belarus, 2021, No 8/37389* (in Russian).

31. On the Development of the Digital Economy. *Decree of the President of the Republic of Belarus, December 21, 2017, No 8 (Amended November 14, 2023, No 357). National Register of Legal Acts of the Republic of Belarus, 2017, No 1/17471* (in Russian).
32. On Approval of the Information Security Concept of the Republic of Belarus. *Resolution of the Security Council of the Republic of Belarus, March 18, 2019, No 1. National Register of Legal Acts of the Republic of Belarus, 2019, No 1/18260* (in Russian).
33. On the State Program "Digital Development of Belarus" for 2021–2025. *Resolution of the Council of Ministers of the Republic of Belarus, February 2, 2021, No 66. National Register of Legal Acts of the Republic of Belarus, 2021, No 5/48748* (in Russian).
34. On Approval of the Rules for the Development of Business Plans for Investment Projects. *Resolution of the Ministry of Economy of the Republic of Belarus, August 31, 2005, No 158 (Amended December 14, 2023, No 25). Minsk: National Center of Legal Information of the Republic of Belarus, 2024* (in Russian).
35. Kim J.-O., Mueller Ch. Y., Klekka Y. R., Oldenderfer M. S., Blashfild R. K. (1989) *Factor, Discriminant, and Cluster Analysis*. Moscow, Finansy i Statistika Publ. (in Russian).
36. Lukasevich I. Ya. (2017) *Investments*. Moscow, Vuzovskiy Uchebnik Publ. (in Russian).
37. Baldin K. V. (2006) *Risk Management*. Moscow, Eksmo Publ. (in Russian).
38. Cattell R. B. (1966) The Scree Test for the Number of Factors. *Multivariate Behavioral Research*. 1 (2), 245–276. DOI: 10.1207/s15327906mbr0102_10.
39. Information Technology – Security Techniques – Information Security Management Systems – Requirements. *ISO/IEC 27001:2022. 3rd ed.* Geneva, ISO/IEC.

Поступила 10.11.2025 | Принята в печать 26.01.2026 | Доступна на сайте 10.04.2026
Received: 10 November 2025 | Accepted: 26 January 2026 | Available on the website: 10 April 2026

Вклад авторов / Authors' contribution

Авторы внесли равный вклад в написание статьи / The authors contributed equally to the writing of the article.

Сведения об авторах

Пискун Е. С., канд. экон. наук, доц. каф. проектирования информационно-компьютерных систем, Белорусский государственный университет информатики и радиоэлектроники (БГУИР)

Азизов А. А., магистрант каф. проектирования информационно-компьютерных систем, БГУИР

Крячев Е. В., магистрант каф. проектирования информационно-компьютерных систем, БГУИР

Information about the authors

Piskun E., Cand. Sci. (Econ.), Associate Professor at the Department of Design Information and Computer Systems, Belarusian State University of Informatics and Radioelectronics (BSUIR)

Azizov A., Master's Student at the Department of Design of Information and Computer Systems, BSUIR

Krychev E., Master's Student at the Department of Design of Information and Computer Systems, BSUIR

Адрес для корреспонденции

220013, Республика Беларусь,
Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 17 292-20-80
E-mail: e.piskun@bsuir.by
Пискун Екатерина Сергеевна

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 17 292-20-80
E-mail: e.piskun@bsuir.by
Piskun Ekaterina