УДК 004.056 Вестник УрФО № 3(57) / 2025, с. 40–50

Нагибин Д. В.

DOI: 10.14529/secur250305

МОДЕЛЬ ОЦЕНКИ ЗРЕЛОСТИ MLOPS С ТОЧКИ ЗРЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ И ОБЪЯСНИМОСТИ

В статье рассматривается модель оценки зрелости MLOps с точки зрения информационной безопасности и объяснимости (XAI). Предлагается структурированный подход к оценке зрелости MLOps, учитывающий специфику XAI, и демонстрируется его потенциальная эффективность в повышении устойчивости моделей к угрозам информационной безопасности. Модель включает критерии оценки, основанные на интеграции XAI в жизненный цикл MLOps, автоматизации анализа угроз, балансе объяснимости и производительности, а также скорости реагирования на угрозы. Представлены уровни зрелости, аналогичные DSOMM OWASP, с примерами применения.

Ключевые слова: *MLOps, информационная безопасность, объяснимый ИИ (XAI), уровни зрелости, безопасность машинного обучения, объяснимость, угрозы информационной безопасности.*

Nagibin D. V.

MLOPS MATURITY MODEL FROM THE PERSPECTIVE OF INFORMATION SECURITY AND EXPLAINABILITY

This article presents a maturity model for MLOps from the perspective of infor-mation security and explainability (XAI). It proposes a structured approach to assessing MLOps maturity, considering the specifics of XAI, and demonstrates its potential effec-tiveness in enhancing the resilience of models against information security threats. The model includes assessment criteria based on the integration of XAI into the MLOps lifecycle, automation of threat analysis, balance of explainability and performance, and speed of response to threats. Maturity levels, similar to DSOMM OWASP, are presented with examples of application.

Keywords: *MLOps, information security, explainable AI (XAI), maturity levels, machine learning security, explainability, information security threats.*

Введение

В современном мире, где системы машинного обучения (МО) все глубже интегрируются в критически важные процессы: от здравоохранения и финансов до автономного транспорта и национальной безопасности – вопросы информационной безопасности (ИБ) приобретают первостепенное значение и требуют комплексного подхода. Однако наряду с потенциальными преимуществами, возрастают и риски информационной безопасности, связанные с уязвимостями, возникающими вследствие использования сложных, зачастую непрозрачных, моделей машинного обучения. Уязвимости, специфичные для систем машинного обучения, такие как отравление данных, атаки на конфиденциальность и эксплуатация уязвимостей в развертывании, представляют серьезную угрозу для надежности и безопасности таких систем [1; 2].

В связи с этим, концепция MLOps становится ключевым элементом обеспечения стабильной и безопасной работы моделей МО на протяжении всего жизненного цикла – от разработки и обучения до развертывания и мониторинга. Однако для достижения максимальной эффективности MLOps, необходимо интегрировать принципы объяснимости искусственного интеллекта (ИИ). Объяснимость ИИ (от англ. explainable artificial intelligence, или XAI) позволяет не только повысить доверие к моделям, но и предоставляет инструменты для выявления и смягчения угроз информационной безопасности, делая системы более устойчивыми [3].

На данный момент уже существует ряд моделей оценки зрелости MLOps (от англ. machine learning operations), таких, как: Microsoft AI Maturity Model, Gartner AI Maturity Model, MITRE AI Framework и OWASP AI Maturity Assessment (AIMA). Но в основном данные модели рассматривают XAI как отдельный инструмент для повышения прозрачности, а не как неотъемлемую часть комплексной модели оценки зрелости. В то же время, интеграция XAI в процессы оценки зрелости MLOps позволяет выявлять и смягчать угрозы, которые не обнаруживаются традиционными методами, а также повышает доверие к моделям и упрощает аудит, что подчеркивает важность интеграции XAI на всех этапах жизненного цикла модели машинного обучения [4].

Таким образом, представляется актуальной разработка комплексной модели оценки зрелости MLOps, интегрированной с принципами объяснимого искусственного интеллекта (XAI), которая позволит систематически оценивать и повышать уровень безопасности и надежности систем машинного обучения.

Данная статья предлагает модель оценки зрелости MLOps с интеграцией принципов объяснимого ИИ (XAI), разработанную для повышения безопасности и интерпретируемости систем машинного обучения. Цель данной работы – предложить структурированный подход к оценке зрелости MLOps, учитывающий специфику XAI, и продемонстрировать его потенциальную эффективность в повышении устойчивости моделей к угрозам информационной безопасности.

Угрозы информационной безопасности в жизненном цикле MLOps

Все более широкое применение систем, использующих машинное обучение, требует разработки новых стратегий обеспечения их безопасности и надежности. MLOps становится ключевым элементом в этом процессе, представляя собой комплексный подход к управлению жизненным циклом моделей машинного обучения.

MLOps включает в себя ряд ключевых аспектов: управление данными, управление версиями моделей, автоматизированное развертывание, непрерывный мониторинг и эффективную интеграцию данных, инфраструктуры и процессов [5]. Однако для достижения максимальной эффективности MLOps, необходимо учитывать вопросы информационной безопасности на каждом этапе жизненного цикла модели МО.

Важными этапами, требующими особого внимания с точки зрения информационной безопасности, являются: управление данными, обучение модели, ее развертывание, мониторинг и обновление. Базовая схема этапов жизненного цикла моделей представлена на рисунке 1.

Применение систем машинного обучения сопряжено с уникальными вызовами к обеспечению информационной безопасности. Например, фаза обучения подвержена риску отравления данных (data poisoning), когда злоумышленник внедряет вредоносные данные для манипулирования результатами обучения. Этап развертывания требует защиты от несанкционированного доступа и

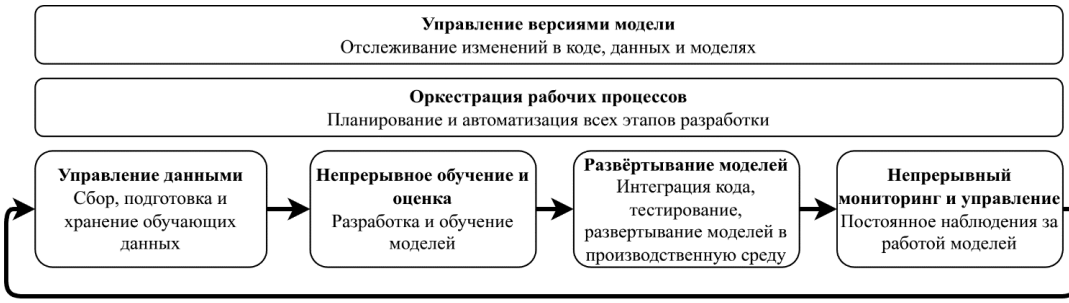

Рис. 1. Базовая схема этапов жизненного цикла MLOps

модификации модели, а также от атак, направленных на эксплуатацию уязвимостей в инфраструктуре. Мониторинг модели должен включать механизмы обнаружения атак, направленных на обход системы защиты, таких как атаки уклонения (evasion attacks). Обновление модели, необходимое для поддержания ее актуальности, должно осуществляться с соблюдением строгих протоколов безопасности, исключающих возможность внедрения вредоносного кода [5].

Также следует отметить атаки, направленные на извлечение конфиденциальной информации из обученных моделей (model extraction), а также манипулирование моделью в процессе ее эксплуатации (adversarial attacks) [2; 6]. Эти атаки могут приводить к непредсказуемым последствиям, включая утечку конфиденциальных данных, нарушение целостности системы и потерю доверия к модели. Особую опасность представляют атаки, эксплуатирующие непрозрачность моделей [7].

Эффективная защита систем машинного обучения требует понимания этих угроз и применения соответствующих мер безопасности на каждом этапе жизненного цикла MLOps, а также активной интеграции методов объяснимого ИИ (XAI) для повышения прозрачности, надежности и верифицируемости моделей.

Объяснимость ИИ как фактор информационной безопасности в MLOps

Объяснимости ИИ (XAI) важна не только для понимания работы моделей, но и для повышения их устойчивости к кибератакам, поскольку она позволяет выявлять потенциальные уязвимости и вредоносные паттерны, например, в «черных ящиках» – моделях, решения которых трудно понять и интерпретировать. В частности, интеграция XAI позволяет выявлять предвзятости в данных и моделях, а также прогнозировать поведение модели в различных сценариях, что критически важно для обеспечения безопасности и надежности систем МО.

Прямое влияние объяснимости на защиту от кибератак заключается в возможности более глубокого анализа моделей. Модели, решения которых можно понять и интерпретировать, легче анализировать на предмет наличия скрытых уязвимостей и предвзятостей. Объяснимость позволяют выявлять потенциальные точки отказа и предсказывать поведение модели в различных сценариях, что существенно повышает устойчивость модели к различным типам атак. Например, понимание того, какие признаки оказывают наибольшее влияние на принятие решения, позволяет выявлять признаки, которые могут быть использованы злоумышленниками для манипулирования моделью [3]. Такой анализ позволяет не только выявлять уязвимости, но и разрабатывать контрмеры для их предотвращения.

Значение объяснимости ИИ трудно переоценить. Например, в медицине объяснимость позволяет врачам понимать, почему модель поставила определенный диагноз или рекомендовала конкретное лечение, что повышает доверие к системе и позволяет принимать более взвешенные решения. В финансах объяснимость может быть полезна для обеспечения прозрачности при принятии решений о кредитовании, страховании и инвестициях [8].

Для эффективного применения объяснимости ИИ в целях защиты моделей, важно понимать различные категории методов и их специфические возможности. Методы объяснимости можно разделить на три основные категории: встроенные (intrinsic), постхок (post-hoc) и априорные (prior) [9].

Априорные (prior) методы разрабатываются с учетом объяснимости изначально, ча-

сто используя простые, линейные модели, которые по своей природе легко интерпретируются. Такие модели могут быть менее точными, но прозрачность делает их идеальными для понимания процесса принятия решений. Например, использование линейной регрессии вместо глубокой нейронной сети для прогнозирования кредитного риска является примером априорного подхода.

Встроенные (intrinsic) методы интегрируются непосредственно в архитектуру модели для обеспечения объяснимость на каждом этапе обучения. Например, механизм внимания (attention mechanism) в нейронных сетях позволяет понять, какие части входных данных наиболее важны для принятия решения [10].

Постхок (post-hoc) методы применяются к уже обученной модели для получения объяснений ее поведения. Они не требуют изменений в архитектуре модели и могут быть применены к любой модели, независимо от ее сложности. Постхок (post-hoc) методы являются наиболее распространенными и гибкими, но могут быть менее точными, чем встроенные (intrinsic) методы [5].

Важно отметить, что объяснимость может быть локальной или глобальной. Локальная объяснимость позволяет понять, почему модель приняла конкретное решение для конкретного экземпляра данных, в то время как глобальная объяснимость позволяет понять общие закономерности, которые использует модель для принятия решений на всем наборе данных [8].

Существует широкий спектр методов объяснимости ИИ, каждый из которых имеет свои особенности и области применения. Рассмотрим некоторые из них.

SHAP (SHapley Additive exPlanations) позволяет достигать как локальной, так и глобальной объяснимости, вычисляя вклад каждого признака в принятие решения [3]. LIME (Local Interpretable Model-agnostic Explanations), в свою очередь, создает локальные, интерпретируемые модели вокруг конкретного предсказания, позволяя понять, какие признаки наиболее важны для данного конкретного случая [8]. What-If Tool позволяет интерактивно исследовать поведение модели и визуализировать влияние различных факторов на результат [11]. ELI5 (Explain Like I'm 5) предоставляет простые и понятные объяснения работы моделей, особенно полезные для неспециалистов [12].

Объяснимость помогает выявлять уязвимости моделей, которые могут быть использованы для атак. Выбор метода объяснимости должен основываться на конкретных требованиях задачи, доступных ресурсах и необходимом уровне детализации. Например, SHAP может использоваться для обнаружения предвзятости, а LIME – для объяснения конкретных предсказаний [3; 8].

Интеграция XAI в MLOps-инфраструктуру

Интеграция методов объяснимого искусственного интеллекта (XAI) в инфраструктуру MLOps требует комплексной работы на каждом этапе жизненного цикла MLOps. Для достижения этой цели необходимо автоматизировать процессы безопасности и использовать специализированные инструменты, позволяющие отслеживать объяснимость моделей, валидировать их поведение и автоматизировать рутинные задачи.

Интеграция объяснимости (XAI) в процессы MLOps находит опору в ряде нормативных документов и инициатив. Так в стандарте NIST IR 8312 выделяются четыре принципа объяснимого ИИ: объяснимость, точность, ограниченность и согласованность [13]. Стандарты ISO/IEC 27001 и сопутствующие из серии 27000, а также ГОСТ Р ИСО/МЭК 27001, устанавливают более общие требования к системам управления информационной безопасностью, акцентируя внимание на объяснимости (XAI).

Ключевым элементом успешной интеграции объяснимости (XAI) является её внедрение в конвейеры непрерывной интеграции и доставки (CI/CD). CI/CD позволяет автоматизировать проверку моделей на предмет объяснимости и безопасности перед развертыванием. Например, при помощи SHAP можно оценивать важность признаков и выявлять потенциальные предвзятости, останавливать развертывание при выходе результатов анализа за допустимые границы, а также выявлять признаки, непропорционально влияющие на предсказания для определенных групп пользователей, что позволяет своевременно устранять предвзятости и повышать доверие к моделям ИИ [5].

Кроме прочего, важно обеспечить постоянный мониторинг моделей в процессе эксплуатации. Изменение статистических свойств входных данных, используемых для обучения модели, – явление, известное как

дрейф данных (data drift), может привести к снижению производительности и увеличению вероятности ошибок [11]. Например, LIME может использоваться для анализа локального поведения модели и выявлять признаки, которые внесли наибольший вклад в предсказание для конкретного экземпляра данных. Мониторинг LIME-анализа может быть полезен для своевременного выявления изменений в важности признаков и сигнализировать о потенциальном дрейфе данных, требующем переобучения модели или корректировки входных данных [14].

Для эффективной интеграции объяснимого искусственного интеллекта (XAI) в процессы MLOps необходим специализированный инструментарий. Платформы, такие как MLflow, позволяют управлять жизненным циклом моделей машинного обучения, включая отслеживание и анализ объяснимости моделей при интеграции с инструментами XAI, например, SHAP. Kubeflow Pipelines, в свою очередь, позволяют автоматизировать проверку моделей на предмет объяснимости и безопасности [11]. Сочетание этих инструментов позволяет не только отслеживать, но и активно управлять объяснимостью моделей в рамках MLOps.

Одним из ключевых вызовов при внедрении XAI в MLOps является поиск оптимального баланса между объяснимостью модели и её точностью. В некоторых случаях, использование более простых, но интерпретируемых моделей, может оказаться предпочтительнее, чем сложные нейронные сети. Данный выбор должен основываться на детальном анализе специфики решаемой задачи и требований к интерпретируемости результатов [15]. Кроме того, необходимо учитывать экономические аспекты внедрения XAI, включая потребность в дополнительных вычислительных ресурсах и временных затратах, связанных с обучением персонала.

Для успешной интеграции XAI рекомендуется придерживаться поэтапного подхода, начиная с оценки текущего уровня зрелости MLOps и определения приоритетных областей для внедрения XAI. Использование гибридных подходов, таких как применение локальной объяснимости для «черных ящиков», может помочь сбалансировать требования к объяснимости и производительности, обеспечивая при этом возможность анализа сложных моделей.

Предлагаемая модель оценки зрелости MLOps с точки зрения информационной безопасности и объяснимости

Интеграция XAI в MLOps может быть формализована как система управления рисками, где каждый уровень зрелости соответствует определенному этапу снижения рисков, характеризующемуся переходом от первого уровня (реактивного), ориентированного на устранение последствий угроз, к проактивному уровню, направленному на предотвращение угроз.

Кроме того, перспективным подходом является рассмотрение MLOps как киберфизической системы. Киберфизические системы (от англ. cyber-physical systems, или CPS) представляют собой интеграцию вычислительных процессов и физических процессов, обеспечивая взаимодействие между программным обеспечением (ПО) и оборудованием. Рассматривая MLOps в рамках CPS, XAI позволяет лучше понимать, как работают модели машинного обучения и почему они принимают те или иные решения, что повышает доверие к моделям и облегчает их использование в критически важных приложениях. Автоматизация безопасности, в свою очередь, снижает вероятность ошибок, которые могут быть вызваны человеческим фактором, например, неправильной настройкой параметров или несоблюдением процедур безопасности [16].

Оценка зрелости MLOps с точки зрения ИБ и объяснимости основывается на четырех ключевых критериях: степень интеграции XAI в этапы жизненного цикла модели (обучение, развертывание, мониторинг), уровень автоматизации анализа угроз (переход от реактивного к предиктивному подходу), баланс между объяснимостью и производительностью (выбор методов, соответствующих требованиям задачи) и скорость реагирования на угрозы с использованием инструментов XAI.

Предлагаемая модель зрелости основана на принципах OWASP's DSOMM [17], адаптированных под специфику MLOps. Цель создания модели – повышение уровня безопасности и надежности систем машинного обучения посредством интеграции XAI в процессы MLOps.

Модель фокусируется на интеграции XAI на всех этапах жизненного цикла моделей в рамках процессов MLOps с обеспечением баланса между объяснимостью и производительностью, а также автоматизации анализа

Таблица 1

Уровни зрелости предлагаемой модели с точки зрения ИБ и XAI

Уровень зрелости	Описание	Интеграция XAI	Автоматизация анализа угроз	Скорость реагирования на угрозы
1. Initial/Ad-hoc (Начальный)	Отсутствие формализованных процедур и структурированного подхода	Отсутствует или минимальная	Реактивный подход (реагирование на инциденты после их возникновения)	Низкая
2. Repeatable (Повторяемый)	Начало структурирования процессов, документирование базовых процедур, внедрение простых инструментов автоматизации	XAI рассматривается как дополнительная функция	Базовый мониторинг	Низкая
3. Defined (Определенный)	Формализованные процессы, внедрение стандартизированных процедур, интеграция XAI в CI/CD-конвейеры	Начальная интеграция объяснимости на отдельных этапах жизненного цикла	Начальный переход к автоматизации	Средняя
4. Managed (Управляемый)	Проактивный подход к управлению рисками, автоматизированный анализ угроз, баланс объяснимости и производительности	Базовая интеграция в процессы разработки и эксплуатации моделей	Предиктивный подход (прогнозирование атак)	Высокая
5. Optimized (Оптимизированный)	Постоянное улучшение процессов, использование AI для выявления и предотвращения угроз, непрерывный мониторинг и адаптация	Интегрирована во все этапы жизненного цикла моделей	Проактивный подход (автоматическое исправление уязвимостей и предотвращение атак)	Очень высокая

угроз с использованием инструментов XAI. Модель включает пять уровней, отражающих прогресс в интеграции XAI и повышении уровня безопасности – таблица 1.

Уровень 1 (initial/ad-hoc – начальный) представляет собой начальный этап развития MLOps, характеризующийся ручным управлением, отсутствием формализованных процедур и минимальной автоматизацией. Обеспечение безопасности строится на основе реактивных мер, а объяснимость не является приоритетом.

Уровень 2 (repeatable – повторяемый) – это этап, на котором процессы MLOps начинают автоматизироваться, появляются базовые процедуры мониторинга и управления версиями моделей. Безопасность обеспечивается на основе стандартных процедур, но отсутствует активное обнаружение атак. Объяснимость рассматривается как дополнительная функция.

Уровень 3 (defined – определенный) представляет собой этап стандартизации и документирования процессов MLOps, охватывающий основные этапы жизненного цикла моделей. Внедряются базовые механизмы мониторинга и обнаружения атак, начинается интеграция методов объяснимости для анализа и интерпретации моделей.

Уровень 4 (managed – управляемый) – это этап, на котором процессы MLOps подвергаются количественной оценке и оптимизации. Внедряются продвинутые механизмы мониторинга и обнаружения атак, и методы

объяснимости интегрированы в процессы разработки и эксплуатации моделей, обеспечивая прозрачность и подотчетность.

Уровень 5 (optimized – оптимизированный) – это высший уровень развития MLOps, характеризующийся непрерывным улучшением процессов на основе данных и обратной связи. Внедряются проактивные меры безопасности и автоматическое исправление уязвимостей, а методы объяснимости интегрированы во все этапы жизненного цикла моделей, обеспечивая максимальную прозрачность и доверие.

В качестве примера можно привести организацию, которая планирует внедрить практики MLOps для автоматизации процессов разработки и развертывания моделей машинного обучения.

Организация начинает свой путь с уровня 1 (initial/ad-hoc – начальный). На данном этапе MLOps существует в виде ряда ручных процессов, связанных с обучением и развертыванием моделей, но без формализованной политики безопасности или объяснимости. Этапы MLOps включают: ручную подготовку данных, обучение моделей с использованием локальных вычислительных ресурсов, ручное тестирование моделей и развертывание моделей в производственную среду. Отсутствует мониторинг объяснимости предсказаний модели.

По мере развития, организация переходит на уровень 2 (repeatable – повторяемый). Разработчики начинают проводить ручное тестирование моделей на предмет предвзятости, используя простые метрики. Начинается сбор данных о предсказаниях модели, но без автоматизированного анализа. Появляются первые попытки автоматизации процессов MLOps. Начинают использоваться инструменты визуализации важности признаков, но анализ объяснимости проводится вручную.

На уровне 3 (defined – определенный) организация разрабатывает и внедряет политику безопасности для моделей MLOps, определяющую требования к тестированию и мониторингу. Процессы MLOps документируются, что позволяет обеспечить воспроизводимость и отслеживаемость. Внедряются базовые инструменты для автоматизации сборки и развертывания моделей, например, использование CI/CD-конвейера для автоматизации развертывания моделей. Инструменты для визуализации важности признаков начинают использоваться систематически, но анализ объяснимости пока в основном проводится вручную.

Переходя на уровень 4 (managed – управляемый), организация интегрирует SHAP в CI/CD-конвейер для автоматической проверки объяснимости моделей перед развертыванием. Используется LIME для мониторинга дрейфа данных и выявления потенциальных угроз безопасности. Разработаны автоматизированные сценарии отката модели в случае обнаружения аномалий. Активно применяются инструменты MLOps.

Наконец, организация достигает уровня 5 (optimized – оптимизированный), что позволяет перейти к проактивным методам защиты от потенциальных угроз безопасности, например на основе анализа исторических данных. Постоянно оптимизируется процессы обеспечения безопасности и объяснимости. Инструменты MLOps позволяют автоматизировать все этапы жизненного цикла модели, включая сбор данных, обучение, развертывание, мониторинг и переобучение. Внедряется автоматизированный анализ причин возникновения аномалий в работе моделей на основе объяснимости.

Для количественной оценки уровня развития практик информационной безопасности и объяснимости в жизненном цикле машинного обучения, можно использовать индекс зрелости MLOps (Maturity Index, MI), вычисляемый например так (1):

$$MI = (A \cdot \alpha) + (B \cdot \beta) + (C \cdot \gamma) + (D \cdot \delta), \quad (1)$$

где A – степень интеграции XAI в этапы жизненного цикла модели;

B – уровень автоматизации анализа угроз;

C – баланс между объяснимостью и производительностью;

D – скорость реагирования на угрозы с использованием инструментов XAI;

коэффициенты α, β, γ и δ отражают относительную важность каждого критерия в конкретном контексте (α для A, β для B, γ для C и δ для D), позволяя адаптировать оценку зрелости к специфическим требованиям и приоритетам. Коэффициенты между собой связаны и влияют на друг друга, поэтому их сумма должна равняться единице, обеспечивая получение осмысленного значения индекса зрелости.

Степень интеграции XAI в этапы жизненного цикла модели может быть определена как:

$$A = \frac{\text{Количество этапов } MLOps \text{ с } XAI}{\text{Общее количество этапов } MLOps},$$

уровень автоматизации анализа угроз:

$$B = \frac{\text{Число автоматизированных процессов}}{\text{Общее число процессов}},$$

баланс между объяснимостью и производительностью:

$$C = \frac{\text{Точность}}{\text{Объяснимость}},$$

скорость реагирования на угрозы с использованием инструментов XAI:

$$D = \frac{1}{T_{\text{реакции}}}.$$

Приведём примерный расчёт индекса зрелости MLOps для организации, описанной в примере выше, демонстрирующий эволюцию её практик информационной безопасности и объяснимости – таблица 2.

Представленная таблица иллюстрирует, как индекс зрелости (MI) увеличивается с каждым уровнем, отражая прогресс организации в области MLOps. В результате применения предложенной модели, организация получает возможность не только объективно оценить текущий уровень информационной безопасности и объяснимости своих моделей, но и спланировать дальнейшие шаги по оптимизации процессов, снижению рисков и повышению общей производительности MLOps-инфраструктуры.

Важно отметить, что определение коэффициентов и оценок должно проводиться экспертами, обладающими знаниями в области MLOps, XAI и информационной безопасности, а формулы и расчёты по ним представлены исключительно для примера, демонстрирующего потенциальное применение предложенной модели. Реальные значения будут зависеть от специфики организации, её приоритетов и используемых технологий.

Заключение

Проведенное исследование продемонстрировало, что интеграция принципов MLOps, подкрепленная активным использованием методов объяснимого искусственного интеллекта (XAI), может оказаться полезной в процессе создания надежных, безопасных и прозрачных моделей ИИ.

Предложенная модель уровней зрелости MLOps предоставляет структурированный подход к оценке и улучшению процессов разработки и эксплуатации моделей машинного обучения. Переход от начального уровня, характеризующегося ручным управлением и отсутствием формализованных процедур, к оптимизированному уровню, обеспечивающему проактивные меры безопасности и автоматическое исправление уязвимостей, требует целенаправленных усилий и инвестиций.

С практической точки зрения, модель может найти применение в различных отраслях, включая финансовый сектор, здравоохранение и государственное управление, где безопасность и надежность моделей ИИ критически важны.

С научной точки зрения, модель вносит вклад в развитие методологии оценки зрелости процессов разработки и эксплуатации ИИ, предлагая интегрированный подход, учитывающий как технические аспекты безопасности, так и принципы объяснимости.

Таблица 2

Примерный расчёт индекса зрелости по формуле (1)

Уровень Зрелости	A	B	C	D	α	β	γ	δ	MI
1. Initial/Ad-hoc – начальный	0	0	0	0	0,3	0,25	0,25	0,2	0
2. Repeatable – повторяемый	0,25	0,25	0,25	0	0,3	0,25	0,25	0,2	0,2
3. Defined – определенный	0,5	0,5	0,5	0,25	0,3	0,25	0,25	0,2	0,45
4. Managed – управляемый	0,75	0,75	0,75	0,5	0,3	0,25	0,25	0,2	0,7
5. Optimized – оптимизированный	1	1	1	1	0,3	0,25	0,25	0,2	1

Предложенная модель способствует решению задач, связанных с повышением доверия к моделям ИИ, снижением рисков кибератак и манипуляций, а также обеспечением соответствия нормативным требованиям. При этом, важно помнить о необходимости достижения оптимального баланса между объяснимостью и сложностью моделей, выбирая стратегии, позволяющие максимизировать точность и прозрачность без существенной потери производительности.

Дальнейшие исследования могут быть направлены на разработку новых методов объяснимости, адаптированных к специфическим требованиям различных областей применения, а также на создание автоматизированных инструментов для оценки и улучшения зрелости MLOps. Предлагаемая модель может оказаться полезной и эффективной при оценке и повышении уровня безопасности и надежности систем МО, однако ее практическая ценность требует дальнейшей верификации и адаптации к конкретным условиям и требованиям.

Литература

1. Намиот Д.Е. О работе AI Red Team / Д.Е. Намиот, Е.В. Зубарева // International Journal of Open Information Technologies. – 2023. – Т. 11. – № 10. – С. 130-139.

2. Намиот Д.Е. Схемы атак на модели машинного обучения / Д.Е. Намиот // International Journal of Open Information Technologies. – 2023. – Vol. 11. – № 5. – P. 68-86.

3. Explainable artificial intelligence for cybersecurity: a literature survey / F. Charmet [et al.] // Annals of Telecommunications. – 2022. – Vol. 77. – Explainable artificial intelligence for cybersecurity. – № 11-12. – P. 789-812.

4. A MLOps Architecture for XAI in Industrial Applications / L. Faubel [et al.] // 2024 IEEE 29th International Conference on Emerging Technologies and Factory Au-tomation (ETFA). – 2024. – P. 1-4.

5. Kreuzberger D. Machine Learning Operations (MLOps): Overview, Definition, and Architecture / D. Kreuzberger, N. Kühl, S. Hirschl // IEEE Access. – 2023. – Vol. 11. – Machine Learning Operations (MLOps). – P. 31866-31879.

6. Намиот Д.Е. Атаки на системы машинного обучения - общие проблемы и методы / Д.Е. Намиот, Е.А. Ильюшин, И.В. Чижов // International Journal of Open Information Technologies. – 2022. – Vol. 10. – № 3. – P. 17-22.

7. Saputra A. The Robustness of Machine Learning Models Using MLSecOps: A Case Study On Delivery Service Forecasting / A. Saputra, E. Suryani, N.A. Rakhmawati // 2023 14th International Conference on Information & Communication Technology and System (ICTS). – 2023. – The Robustness of Machine Learning Models Using MLSecOps. – P. 265-270.

8. Шевская Н.В. Объяснимый искусственный интеллект и методы интерпретации результатов / Н.В. Шевская // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 2(33). – С. 24-25.

9. Explainable Artificial Intelligence (XAI): What we know and what is left to at-tain Trustworthy Artificial Intelligence / S. Ali [et al.] // Information Fusion. – 2023. – Vol. 99. – Explainable Artificial Intelligence (XAI). – P. 101805.

10. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes / D. Soydaner // Neural Computing and Applications. – 2022. – Vol. 34. – Attention mechanism in neural networks. – № 16. – P. 13371-13385.

11. Mona M. Google Cloud Certified Professional Machine Learning Engineer Study Guide / M. Mona, P. Ramamurthy. – Indianapolis: John Wiley and Sons, 2023.

12. Мишра П. Объяснимые модели искусственного интеллекта на Python. Модель искусственного интеллекта. Объяснения с использованием библиотек, расширений и фреймворков на основе языка Python. Объяснимые модели искусственного интеллекта на Python / П. Мишра. – ДМК Пресс. – Москва: Apress, 2022. – 344 с.

13. Four principles of explainable artificial intelligence / P.J. Phillips [et al.]. – Gaithersburg, MD: National Institute of Standards and Technology (U.S.), 2021.

14. Digital Image Security: techniques and applications. Digital Image Security / eds. A.K. Singh [et al.]. – S.l.: CRC PRESS, 2024. – 351 p.

15. Investigating the Duality of Interpretability and Explainability in Machine Learning / M. Garouani [и др.] // 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). – 2024. – С. 861-867.

16. MLOps for Cyberphysical Production Systems: Challenges and Solutions / L. Faubel [и др.] // IEEE Software. – 2025. – Т. 42. – MLOps for Cyberphysical Production Systems. – № 1. – С. 65-73.

17. OWASP DevSecOps Maturity Model [Электронный ресурс]. – URL: https://owasp.org/www-project-devsecops-maturity-model/ (дата обращения: 31.03.2025).

References

1. Namiot D.Ye. O rabote AI Red Team / D.Ye. Namiot, Ye.V. Zubareva // In-ternational Journal of Open Information Technologies. – 2023. – T. 11. – № 10. – S. 130-139.

2. Namiot D.Ye. Skhemy atak na modeli mashinnogo obucheniya / D.Ye. Namiot // International Journal of Open Information Technologies. – 2023. – Vol. 11. – № 5. – P. 68-86.

3. Explainable artificial intelligence for cybersecurity: a literature survey / F. Charmet [et al.] // Annals of Telecommunications. – 2022. – Vol. 77. – Explainable artificial intelligence for cybersecurity. – № 11-12. – P. 789-812.

4. A MLOps Architecture for XAI in Industrial Applications / L. Faubel [et al.] // 2024 IEEE 29th International Conference on Emerging Technologies and Factory Au-tomation (ETFA). – 2024. – P. 1-4.

5. Kreuzberger D. Machine Learning Operations (MLOps): Overview, Definition, and Architecture / D. Kreuzberger, N. Kühl, S. Hirschl // IEEE Access. – 2023. – Vol. 11. – Machine Learning Operations (MLOps). – P. 31866-31879.

6. Namiot D.Ye. Ataki na sistemy mashinnogo obucheniya - obshchiye problemy i metody / D.Ye. Namiot, Ye.A. Il'yushin, I.V. Chizhov // International Journal of Open Information Technologies. – 2022. – Vol. 10. – № 3. – P. 17-22.

7. Saputra A. The Robustness of Machine Learning Models Using MLSecOps: A Case Study On Delivery Service Forecasting / A. Saputra, E. Suryani, N.A. Rakhmawati // 2023 14th International Conference on Information & Communication Technology and System (ICTS). – 2023. – The Robustness of Machine Learning Models Using MLSecOps. – P. 265-270.

8. Shevskaya N.V. Ob"yasnimyy iskusstvennyy intellekt i metody inter-pretatsii rezul'tatov / N.V. Shevskaya // Modelirovaniye, optimizatsiya i infor-matsionnyye tekhnologii. – 2021. – T. 9. – № 2(33). – S. 24-25.

9. Explainable Artificial Intelligence (XAI): What we know and what is left to at-tain Trustworthy Artificial Intelligence / S. Ali [et al.] // Information Fusion. – 2023. – Vol. 99. – Explainable Artificial Intelligence (XAI). – P. 101805.

10. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes / D. Soydaner // Neural Computing and Applications. – 2022. – Vol. 34. – Attention mechanism in neural networks. – № 16. – P. 13371-13385.

11. Mona M. Google Cloud Certified Professional Machine Learning Engineer Study Guide / M. Mona, P. Ramamurthy. – Indianapolis: John Wiley and Sons, 2023.

12. Mishra P. Ob"yasnimyye modeli iskusstvennogo intellekta na Python. Model' iskusstvennogo intellekta. Ob"yasneniya s ispol'zovaniyem bibliotek, rasshireniy i freymvorkov na osnove yazyka Python. Ob"yasnimyye modeli is-kusstvennogo intellekta na Python / P. Mishra. – DMK Press. – Moskva: Apress, 2022. – 344 s.

13. Four principles of explainable artificial intelligence / P.J. Phillips [et al.]. – Gaithersburg, MD: National Institute of Standards and Technology (U.S.), 2021.

14. Digital Image Security: techniques and applications. Digital Image Security / eds. A.K. Singh [et al.]. – S.l.: CRC PRESS, 2024. – 351 p.

15. Investigating the Duality of Interpretability and Explainability in Machine Learning / M. Garouani [и др.] // 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). – 2024. – С. 861-867.

16. MLOps for Cyberphysical Production Systems: Challenges and Solutions / L. Faubel [и др.] // IEEE Software. – 2025. – Т. 42. – MLOps for Cyberphysical Production Systems. – № 1. – С. 65-73.

17. OWASP DevSecOps Maturity Model [Электронный ресурс]. – URL: https://owasp.org/www-project-devsecops-maturity-model/ (дата обращения: 31.03.2025).

НАГИБИН Дмитрий Викторович, аспирант кафедры «Автоматика и телемеханика» федерального государственного автономного образовательного учреждения высшего образования «Пермский национальный исследовательский политехнический университет». 614990, г. Пермь, Комсомольский проспект, д. 29. E-mail: dvnagibin@pstu.ru

NAGIBIN Dmitriy Viktorovich, post-graduate student of the Department of «Automation and Telemechanics» of the Federal Autonomous Educational Institution of Higher Education «Perm National Research Polytechnic University». 614990, Perm, Komsomolsky prospekt, 29. E-mail: dvnagibin@pstu.ru