УДК 004.94
МРТИ 28.23.15
DOI 10.56525/DGKT1498

ИНТЕГРАЦИЯ ОБЪЯСНИМОГО ИИ (XAI) В ЗАЩИЩЕННЫЕ КОНТУРЫ УПРАВЛЕНИЯ СИСТЕМАМИ

[1]Жилкишбаева Г.С.*, **[1]Утебай О.Қ., [2]Gadirli A.V., [1]Қожабай Қ.Б.**
[1]Университет Есенова, г.Ақтау. Казакстан
[2]Азербайджанский государственный университет нефти и промышленности, Баку, Азербайджан
e-mail: ytebai.oralbay@yu.edu.kz, anargadirli55@gmail.com, gulnaz.zhilkishbayeva@yu.edu.kz, kairatkozhabay@gmail.com

Аннотация. Внедрение алгоритмов машинного обучения и глубоких нейронных сетей в критически важные киберфизические системы (CPS) требует перехода от стохастических моделей «черного ящика» к прозрачным архитектурным решениям, обеспечивающим детерминированную безопасность. Техническая необходимость интеграции объяснимого искусственного интеллекта (XAI) продиктована требованиями верификации и валидации (V&V) в недетерминированных средах, где цена ошибки исключает использование непроверяемых эвристик. Данная работа исследует фундаментальный конфликт между предсказательной мощностью моделей и их интерпретируемостью в реальном времени. Рассматриваются таксономии XAI, включая внутренние (intrinsic) и постфактумные (post-hoc) методы, а также их влияние на задержки в контурах управления (control loops). Анализируются архитектурные компромиссы при использовании аппаратных ускорителей на базе FPGA и специализированных ASIC. Особое внимание уделяется устойчивости модулей объяснения к адверсариальным воздействиям, таким как «отмывание справедливости» (fairwashing) и манипуляция признаками. Синтезируются подходы к созданию адаптивных систем объяснения на основе неявной обратной связи от оператора. В заключении выделяются критические технологические пробелы в области стандартизации метрик интерпретируемости и обеспечения безопасности самих алгоритмов XAI [1-3].

Ключевые слова. Объяснимый ИИ (XAI), киберфизические системы (CPS), задержка (latency), адверсариальное машинное обучение, детерминированная безопасность, контуры управления, верификация моделей, глубокое обучение с подкреплением (DRL).

Введение

Современная инфраструктура управления критическими объектами — от атомных энергетических установок до автономных транспортных сетей — претерпевает радикальную трансформацию, связанную с интеграцией интеллектуальных агентов на базе глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL). В отличие от традиционных контроллеров, таких как пропорционально-интегрально-дифференцирующие регуляторы (PID) или системы прогностического управления моделями (MPC), нейросетевые модели способны адаптироваться к сложным, динамически меняющимся нелинейным средам. Однако высокая предсказательная способность этих моделей сопряжена с критическим недостатком — отсутствием прозрачности внутренних механизмов принятия решений. В защищенных контурах управления, где время принятия решения исчисляется миллисекундами, использование «черных ящиков» создает неприемлемые риски, связанные с возможным смещением (bias), катастрофическим забыванием и непредсказуемым поведением в граничных сценариях (edge cases) [1-2, 4].

Конфликт между производительностью и проверяемостью становится основным барьером для лицензирования ИИ-систем регуляторами, такими как Комиссия по ядерному регулированию США (NRC). Традиционные методы верификации программного обеспечения, основанные на статическом анализе и полном покрытии путей выполнения, не применимы к архитектурам с миллиардами обучаемых параметров. Интеграция методов XAI призвана устранить этот разрыв, предоставляя математически обоснованные объяснения причинно-следственных связей между входными векторами сенсоров и выходными управляющими воздействиями. Технологическая необходимость XAI в защищенных контурах заключается в обеспечении «прослеживаемости» (traceability) и возможности аудита решений в реальном времени [1-2, 5].

Тем не менее, внедрение слоев объяснимости в архитектуру CPS порождает ряд инженерных компромиссов. Основной из них — противоречие между глубиной интерпретации и вычислительной задержкой (latency). Алгоритмы, такие как SHAP (SHapley Additive exPlanations) или LIME (Local Interpretable Model-agnostic Explanations), требуют значительных ресурсов CPU/GPU, что может привести к нарушению жестких временных регламентов реального времени. Кроме того, сами модули объяснения становятся новой поверхностью атаки: злоумышленники могут манипулировать объяснениями, чтобы скрыть вредоносное вмешательство в работу основной модели или дезинформировать оператора системы. Таким образом, проектирование систем с XAI требует комплексного подхода, сочетающего формальную математическую логику, аппаратную оптимизацию и криптографическую защиту целостности объяснений [3, 6-7].

Материалы и методы. В данной работе используется систематический подход к анализу фреймворков XAI, интегрированных в киберфизические контуры управления. Источниками данных послужили публикации ведущих научно-исследовательских институтов (IEEE, ACM), а также технические спецификации промышленных стандартов безопасности (ISO 26262, IEC 62304). Критерии отбора материалов включали наличие экспериментальных данных о задержках, математическую формализацию методов объяснения и применимость в системах с высокими требованиями к надежности (fail-safe) [1, 5, 8].

Для оценки архитектурных решений использовались следующие критерии анализа:

1. Тип интерпретируемости: внутренняя (intrinsic) против постфактумной (post-hoc). Анализировалось, как выбор типа модели влияет на возможность формальной верификации.

2. Область объяснения: локальная против глобальной. Оценивалась значимость локальных объяснений для диагностики единичных инцидентов в CPS против глобальных объяснений для понимания общей стратегии управления.

3. Интеграция в петлю управления (Control Loop): анализировалось место модуля XAI в архитектуре — параллельное выполнение (Observer pattern) или последовательная фильтрация (Gatekeeper pattern).

4. Аппаратная реализация: сравнение эффективности выполнения алгоритмов на базе CPU общего назначения, GPU и специализированных FPGA/ASIC решений [9-11].

Особое внимание уделено синтезу таксономии применения XAI в обратной связи, где объяснение служит не только для информирования человека, но и в качестве входного сигнала для автоматизированных систем безопасности и дообучения моделей. В качестве базового примера рассматривался фреймворк Deterministic Assurance Framework (DTAF), обеспечивающий лицензируемую проверку DRL-агентов в энергетических системах [2, 12-13].

Результаты. Интеграция XAI в защищенные контуры требует понимания математической природы объяснений. В контексте CPS объяснение рассматривается как модель пониженной сложности, аппроксимирующая поведение сложного контроллера в локальной окрестности текущего состояния [13].

Категория метода	Типичные алгоритмы	Математический базис	Преимущества в CPS	Недостатки
Внутренние (Intrinsic)	Decision Trees, GAM, EBM	Прозрачная структура весов/правил	Гарантированный детерминизм	Ограниченная емкость модели
Постфактумные агностические	LIME, SHAP, KernelSHAP	Теория игр, локальная регрессия	Работа с любым контроллером	Высокая задержка, стохастичность
Градиентные (Model-specific)	Grad-CAM, Integrated Gradients	Дифференцирован не по входу	Относительная скорость на GPU	Зависимость от архитектуры NN
Примерные (Example-based)	Counterfactuals (CFE)	Оптимизация близости и целевого класса	Поиск путей вывода из опасного состояния	Сложность поиска в RT

Центральным методом является SHAP, основанный на значениях Шепли из кооперативной теории игр. Для признака j значение $\boldsymbol{\phi}$ определяется как:

$$\phi_j = \sum_{S \subseteq \mathrm{N}\{\mathrm{j}\}}^{n} \frac{(|S|!\,|N| - |S| - 1)!}{|N|!}$$

Этот подход обеспечивает свойство «локальной точности» (local accuracy): сумма вкладов всех признаков в точности равна отклонению прогноза от среднего значения. Это критически важно для аудита безопасности, так как позволяет точно распределить ответственность между сенсорными входами [14-16].

Вторым значимым методом является LIME, минимизирующий функцию потерь для локального суррогата g:

$$\underset{x}{L(f, g, \pi)} + \Omega(g)$$

где π- ядро близости, определяющее локальность вокруг текущего состояния системы x.

В контурах управления LIME позволяет быстро понять, почему контроллер принял решение об аварийном отключении, аппроксимируя сложную нейронную сеть простой линейной моделью в момент инцидента [14, 17].

В системах с защитой от сбоев (fail-safe) время цикла управления часто ограничено интервалом в 1–10 мс. Традиционные реализации KernelSHAP требуют сотен или тысяч вычислений модели для одного объяснения, что делает их непригодными для встраиваемых систем. Исследования показывают, что использование стандартного SHAP может приводить к задержкам до нескольких секунд на сложных многомерных данных [18-19].

Для решения этой проблемы предложены оптимизированные архитектуры, такие как Explainable URLLC Decision Engine (EUDE). В рамках этой системы слой объяснения интегрирован непосредственно в контур управления связью 5G/6G.

Метод	Средняя задержка (мс)	Надежность (%)	Оценка объяснимости
Стандартный DRL	0.94	99.984	0.08
EUDE (Hybrid XAI)	1.03	99.972	0.86

Анализ данных EUDE показывает, что включение XAI увеличивает задержку на 9.6% (0.09 мс). Несмотря на этот рост, система остается в рамках спецификаций сверхнадежной связи с низкой задержкой (URLLC), обеспечивая при этом рост доверия операторов на 52% [7].

Дальнейшее снижение задержек достигается за счет методов амортизации, таких как FastSHAP и SpectralSHAP. FastSHAP обучает отдельную нейронную сеть-объяснитель, которая выдает значения важности признаков за один прямой проход (forward pass). SpectralSHAP использует спектральное разложение Фурье на булевом гиперкубе, что позволяет аналитически вычислять значения Шепли.

Характеристика	KernelSHAP	FastSHAP	SpectralSHAP
Вычислительная сложность	Экспоненциальная $O(2^n)$	Константная (после обучения)	Полиномиальная
Задержка на CPU	3 – 10 с	<100 мс	50 – 200 мс
Задержка на GPU/FPGA	н/д	<1 мс	<5 мс
Требования к памяти	Низкие	Высокие (модель)	Средние

Для систем с критическими требованиями к реальному времени (например, управление квантовыми битами или стабилизация энергосетей) единственным жизнеспособным решением является аппаратная реализация на FPGA. Использование прямого интерфейса между сенсорами (например, EMCCD-камерами) и логикой нейросетевого детектора на FPGA позволяет достичь наносекундных задержек, исключая накладные расходы на шины передачи данных и операционные системы [20, 21].

В защищенных контурах XAI рассматривается как механизм безопасности, однако он сам подвержен адверсариальным атакам. Злоумышленник, обладающий знаниями об алгоритме объяснения, может манипулировать входными данными так, чтобы модель сохраняла свою (возможно, вредоносную) предсказательную логику, но выдавала объяснение, которое выглядит легитимным для аудитора [3, 6].

Выделяются три основных типа атак на XAI в кибербезопасности:

1. Отмывание справедливости (Fairwashed Explanation, FE): Цель атаки — скрыть влияние «чувствительных» или скомпрометированных признаков. Математически это формулируется как поиск возмущения, минимизирующего расстояние между вектором важности признаков и нулевым вектором для защищаемого набора признаков F_S.

2. Манипулируемое объяснение (Manipulated Explanation, ME): Агрессор заставляет систему выдавать произвольно выбранное объяснение g_{target}. Это достигается путем оптимизации малого шума δ, который изменяет градиенты или веса в суррогатной модели объяснителя, не меняя итоговый класс классификации.

3. Атаки с использованием бэкдоров (Backdoor-enabled, BD): В модель внедряется скрытый триггер. В нормальных условиях XAI работает корректно, но при наличии триггера объяснение полностью искажается, скрывая присутствие вредоносного кода или аномалии [3].

Эксперименты на наборах данных для обнаружения фишинга и вредоносного ПО (например, UNSW-NB15) показали, что методы SHAP и LIME крайне чувствительны к таким манипуляциям. В частности, атаки типа ME способны изменить порядок важности топ-5

признаков в 85% случаев без потери точности основной модели. Это создает критический риск: аналитик по безопасности может проигнорировать реальную угрозу, полагаясь на «зеленый» отчет системы объяснения [3, 22].

Для обеспечения детерминированной безопасности в нелинейных системах управления рекомендуется использование паттерна «Привратник» (Gatekeeper). В этой архитектуре модуль XAI работает как активный фильтр между интеллектуальным агентом и исполнительными механизмами [10, 23].

Архитектурная схема «Привратника» включает:

- Слой верификации (Verification Layer): проверяет объяснение на соответствие физическим законам и правилам безопасности. Если SHAP-значение критического датчика указывает на то, что решение принято на основе шума, команда блокируется.
- Детектор выхода за пределы распределения (OOD Detection): использует методы XAI для оценки неопределенности. В системах, таких как ODiSAR (Self-adaptive Robots), цифровой двойник на базе Transformer прогнозирует состояние системы и сравнивает его с реальным. Ошибка реконструкции и предиктивная дисперсия (Monte Carlo dropout) служат индикаторами аномалий [24-25].
- Адаптивная обратная связь (AXTF): Фреймворк Adaptive Explainability Trust Framework (AXTF) динамически настраивает уровень детализации объяснений на основе физиологического состояния оператора (ЭЭГ, ЭКГ, отслеживание взгляда). Если когнитивная нагрузка оператора высока, система выдает краткие визуальные алерты; в спокойном режиме — подробные отчеты о причинно-следственных связях [13].

Таблица ниже иллюстрирует эффективность различных архитектурных подходов к интеграции обратной связи:

Паттерн	Роль XAI	Влияние на задержку	Уровень безопасности
Observer	Пассивный аудит, логирование	Отсутствует	Низкий (пост -фактум)
Gatekeeper	Активная фильтрация воздействий	Высокое (+10-20%)	Высокий (превентивный)
Hybrid	Обучение с учителем в петле	Среднее	Средний (адаптивный)
Deterministic Assurance (DTAF)	Формальная верификация границ	Критическое	Максимальный (лицензируемый)

Фреймворк DTAF заслуживает особого внимания, так как он переводит поведение контроллера в форму доказательства, пригодного для лицензирования. Он использует «детерминированные лицензионные шлюзы», связанные с жесткими пределами безопасности (например, Total Time Unsafe = 0). Агент DRL должен пройти через портфель адверсариальных стресс-тестов, где XAI-прослеживаемость подтверждает, что агент не нарушает границы даже в худших сценариях [2].

Заключение. Интеграция XAI в защищенные контуры управления критическими системами является не факультативным улучшением пользовательского интерфейса, а фундаментальным требованием кибербезопасности и системной инженерии. Проведенный анализ показывает, что современные методы постфактумной интерпретации (SHAP, LIME) сталкиваются с серьезными препятствиями в виде вычислительной задержки и уязвимости к адверсариальным атакам. Однако развитие специализированных аппаратных ускорителей на базе FPGA и внедрение оптимизированных алгоритмов, таких как SpectralSHAP и FastSHAP, позволяют преодолеть барьер реального времени [1, 7, 19].

Критический обзор области выявляет следующие существенные пробелы, требующие дальнейших исследований:

1. Отсутствие стандартизированных метрик «верности объяснения» (explanation fidelity): в настоящее время не существует единого математического критерия, позволяющего оценить, насколько точно XAI отражает истинную логику глубокой нейронной сети в динамических системах.

2. Недостаточная защита модулей XAI: Существующие фреймворки практически не имеют встроенных механизмов противодействия атакам типа «отмывание справедливости» и манипуляции признаками, что делает их потенциальным вектором дезинформации операторов.

3. Разрыв в семантике: Большинство методов XAI предоставляют низкоуровневые коэффициенты важности признаков, которые трудно интерпретировать в терминах высокоуровневых инженерных концепций и физических ограничений. Требуется развитие нейросимволических подходов, объединяющих статистическую мощь нейросетей с логической строгостью экспертных систем [9, 26-27].

Для успешного внедрения ИИ в ядерную энергетику, авиацию и медицину необходимо создание комплексных архитектур, где XAI интегрирован на уровне «Привратника», обеспечивая детерминированные гарантии безопасности в условиях неопределенности. Будущее отрасли лежит в переходе от объяснения «почему модель это сделала» к верификации того, что «модель не сделает ничего недопустимого» [2, 5, 28].

ЛИТЕРАТУРЫ

1. Shankar V. Explainable AI (xAI) in Critical Decision Systems: Algorithm, Frameworks, and Case Studies: дис. … / V. Shankar. – 2022. – DOI: 10.13140/RG.2.2.10504.28169.

2. Abdelrahman Ibrahim A. A Deterministic Assurance Framework for Licensable Explainable AI Grid-Interactive Nuclear Control / A. Abdelrahman Ibrahim, H.-K. Lim // Energies. – 2025. – Vol. 18, № 23. – P. 6268. – DOI: 10.3390/en18236268.

3. Mia M. Explainable but Vulnerable: Adversarial Attacks on XAI Explanation in Cybersecurity Applications / M. Mia, M. M. A. Pritom. – 2025. – arXiv:2510.03623.

4. Yusuf H. U. Architectural Transformations and Emerging Verification Demands in AI-Enabled Cyber-Physical Systems / H. U. Yusuf, K. Gaaloul. – 2025. – arXiv:2510.00519.

5. How to Ensure Safety in AI/ML-Driven Embedded Systems [Электронный ресурс] // Parasoft. – URL: https://www.parasoft.com/white-paper/ensure-safety-ai-embedded-systems/ (дата обращения: 12.01.2026).

6. Baniecki H. Adversarial attacks and defenses in explainable artificial intelligence: A survey / H. Baniecki, P. Biecek // Information Fusion. – 2024. – Vol. 107. – P. 102303.

7. Balogun P. Explainable AI Models for Real-Time Decision-Making in Ultra-Reliable Low-Latency Communications (URLLC) / Peter Balogun, Vangala Murthy. – 2025.

8. Saarela M. Recent applications of Explainable AI (XAI): A systematic literature review / M. Saarela, V. Podgorelec // Applied Sciences. – 2024. – Vol. 14, № 19. – P. 8884.

9. Dib L. Classifying XAI Methods to Resolve Conceptual Ambiguity / L. Dib, L. Capus // Technologies. – 2025. – Vol. 13, № 9. – P. 390.

10. Architecture Patterns for Scaling AI Guardrails [Электронный ресурс] // Galileo. – URL: https://galileo.ai/blog/scaling-ai-guardrails-architecture-patterns (дата обращения: 12.01.2026).

11. Edge AI in Live Production: FPGA / ASIC for Transcoding and Real-Time Analytics [Электронный ресурс] // Promwad. – URL: https://promwad.com/news/edge-ai-live-production-fpga-asic-transcoding-analytics (дата обращения: 12.01.2026).

12. Fernando N. Adaptive XAI in High Stakes Environments: Modeling Swift Trust with

Multimodal Feedback in Human AI Teams / N. Fernando, B. Nakisa, A. Ahmad, M. N. Rastgoo. – 2025. – arXiv:2507.21158.

13. Porcari F. eXplainable AI for data driven control: an inverse optimal control approach / F. Porcari, D. Materassi, S. Formentin. – 2025. – arXiv:2504.11446.

14. Marín Díaz G. Comparative analysis of explainable AI methods for manufacturing defect prediction: A mathematical perspective / G. Marín Díaz // Mathematics. – 2025. – Vol. 13, № 15.– P. 2436.

15. Introduction to SHAP, LIME, and Integrated Gradients [Электронный ресурс] // Patsnap Eureka. – URL:https://eureka.patsnap.com/article/introduction-to-shap-lime-and-integrated-gradients (дата обращения: 12.01.2026).

16. Jethani N. Fastshap: Real-time shapley value estimation / N. Jethani, M. Sudarshan, I. Covert,S. I. Lee, R. Ranganath // ICLR. – 2022.

17. Givisis I. Comparing Explainable AI Models: SHAP, LIME, and Their Role in Electric Field Strength Prediction over Urban Areas / I. Givisis, D. Kalatzis, C. Christakis, Y. Kiouvrekis // Electronics. – 2025. – Vol. 14, № 23. – P. 4766.

18. Olsen L. H. B. Improving the Sampling Strategy in KernelSHAP / L. H. B. Olsen, M. Jullum. – 2024. – arXiv:2410.

19. Onchis D. A Real-Time Capable Universal Explainer for Image Models and Beyond / Darian Onchis. – 2025.

20. Lou B. Low-Latency FPGA Control System for Real-Time Neural Network Processing in CCD-Based Trapped-Ion Qubit Measurement / B. Lou, G. D. Krishnaswaroop, F. Wojcicki [и др.]. – 2025. – arXiv:2512.15838.

21. Agrawal N. AI models on FPGA chips: Advantages and challenges [Электронный ресурс] / Neeraj Agrawal // Medium. – URL:https://medium.com/@neeraj8us/ai-models-on-fpga-chips-advantages-and-challenges-df3955e63 d27 (дата обращения: 12.01.2026).

22. Kuppa A. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security / A. Kuppa, N. A. Le-Khac // 2020 International Joint Conference on Neural Networks (IJCNN). – IEEE, 2020. – P. 1-8.

23. Deterministic AI vs Non-Deterministic AI: Understanding the Core Difference [Электронный ресурс] // Kubiya. – URL: https://www.kubiya.ai/blog/deterministic-ai-vs-non-deterministic-ai (дата обращения: 12.01.2026).

24. Weng C. Out-of-Distribution Detection as a Risk-Control Strategy for Medical Classification Machine Learning Models / C. Weng, J. Ward, W. Lin [и др.] // Clinical and Translational Science. – 2025. – Vol. 18, № 10. – P. e70349.

25. Isaku E. Out of Distribution Detection in Self-adaptive Robots with AI-powered Digital Twins / E. Isaku, H. Sartaj, S. Ali [и др.]. – 2025. – arXiv:2509.12982.

26. Kabir S. A review of explainable artificial intelligence from the perspectives of challenges and opportunities / S. Kabir, M. S. Hossain, K. Andersson // Algorithms. – 2025. – Vol. 18, № 9.– P. 556.

27. hbaniecki/adversarial-explainable-ai: Adversarial attacks on explanations and how to defend them [Электронный ресурс] // GitHub. – URL: https://github.com/hbaniecki/adversarial-explainable-ai (дата обращения: 12.01.2026).

28. Kolapo A. A. Explainable AI in Industrial Cybersecurity: Building Trust in ML-Based Threat Detection Systems / Abdul Azeez Kolapo, Andrew James, Aremu Isaac. – 2024.

ТҮСІНДІРІЛЕТІН AI (XAI) ЖҮЙЕЛЕРДІ БАСҚАРУДЫҢ ҚОРҒАЛҒАН КОНТУРЛАРЫНА ИНТЕГРАЦИЯЛАУ

[1]Жилкишбаева Г.С.*, **[1]Утебай О.Қ.**, **[2]Gadirli A.V.**, **[1]Қожабай Қ.Б.**
[1]Есенов университеті, Ақтау қ. Қазақстан
[2]Әзірбайжан Мемлекеттік Мұнай Және Өнеркәсіп Университеті, Баку, Азербайджан
e-mail: ytebai.oralbay@yu.edu.kz, anargadirli55@gmail.com,
gulnaz.zhilkishbayeva@yu.edu.kz, kairatkozhabay@gmail.com

Аңдатпа. Машиналық оқыту алгоритмдері мен терең нейрондық желілерді маңызды киберфизикалық жүйелерге (CPS) енгізу стохастикалық «қара жәшік» модельдерінен детерминирленген қауіпсіздікті қамтамасыз ететін ашық архитектуралық шешімдерге көшу қажеттігін туындатады. Түсіндірілетін жасанды интеллектіні (XAI) интеграциялаудың техникалық қажеттілігі тексеру және валидация (V&V) талаптарымен байланысты, әсіресе қателіктің құны өте жоғары болатын, сондықтан тексерілмеген эвристикаларды қолдануға болмайтын недетерминирленген орталарда.

Бұл жұмыс модельдердің болжамдық қуаты мен олардың нақты уақыттағы интерпретациялануы арасындағы іргелі қайшылықты зерттейді. XAI таксономиялары қарастырылады, соның ішінде ішкі (intrinsic) және постфактумдық (post-hoc) әдістер, сондай-ақ олардың басқару контурларындағы (control loops) кідірістерге әсері. FPGA негізіндегі аппараттық жеделдеткіштер мен мамандандырылған ASIC қолдану кезіндегі архитектуралық ымыралар талданады.

Ерекше назар түсіндіру модульдерінің адверсариалдық әсерлерге төзімділігіне аударылады, мысалы «әділдікті жуу» (fairwashing) және белгілерді манипуляциялау. Оператордың жасырын кері байланысына негізделген бейімделмелі түсіндіру жүйелерін құру тәсілдері синтезделеді. Қорытындыда интерпретациялану метрикаларын стандарттау және XAI алгоритмдерінің өз қауіпсіздігін қамтамасыз ету салаларындағы негізгі технологиялық олқылықтар анықталады [1–3].

Түйін сөздер: түсіндірілетін жасанды интеллект (XAI), киберфизикалық жүйелер (CPS), кідіріс (latency), адверсариалдық машиналық оқыту, детерминирленген қауіпсіздік, басқару контурлары, модельдерді верификациялау, терең нығайтпалы оқыту (DRL).

INTEGRATION OF EXPLICABLE AI (XAI) INTO THE PROTECTED CONTROL CIRCUITS OF THE SYSTEMS

[1]Zhilkishbaeva G.S.*, **[1]Utebai O.K.**, **[2]Gadirli A.V.**, **[1]Kozhabai K.B.**
[1]Yessenov University, Aktau, Kazakhstan
[2]Azerbaijan State Oil and Industry University, Baku, Azerbaijan
e-mail: ytebai.oralbay@yu.edu.kz, anargadirli55@gmail.com,
gulnaz.zhilkishbayeva@yu.edu.kz, kairatkozhabay@gmail.com

Abstract. The integration of machine learning algorithms and deep neural networks into safety-critical cyber-physical systems (CPS) requires a transition from stochastic “black-box” models to transparent architectural solutions that ensure deterministic safety. The technical necessity of integrating explainable artificial intelligence (XAI) is driven by verification and validation (V&V) requirements in non-deterministic environments where the cost of error precludes the use of unverified heuristics.

This work investigates the fundamental conflict between the predictive power of models and their real-time interpretability. Taxonomies of XAI are examined, including intrinsic and post-hoc methods, as well as their impact on latency in control loops. Architectural trade-offs associated with the use of hardware accelerators based on FPGA and specialized ASICs are analyzed.

Particular attention is given to the robustness of explanation modules against adversarial influences such as fairwashing and feature manipulation. Approaches to developing adaptive explanation systems based on implicit operator feedback are synthesized. The paper concludes by identifying critical technological gaps in the standardization of interpretability metrics and in ensuring the security of XAI algorithms themselves [1–3].

Keywords: Explainable Artificial Intelligence (XAI), Cyber-Physical Systems (CPS), latency, adversarial machine learning, deterministic safety, control loops, model verification, deep reinforcement learning (DRL).