

ОБЪЯСНИМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (ХАИ): ОБЕСПЕЧЕНИЕ ПРОЗРАЧНОСТИ И ДОВЕРИЯ ПОЛЬЗОВАТЕЛЕЙ В ОБРАЗОВАТЕЛЬНЫХ РЕШЕНИЯХ — КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ

Кенджаева Дилдора Худайбергеновна
Ташкентский университет прикладных наук,
старший преподаватель кафедры «Компьютерный инжиниринг»
E-mail: shainaxon@gmail.com | Google Scholar: scholar.google.com/citations?user=c0FaPIAAAAJ

Аннотация: В статье исследована проблема применения методов объяснимого искусственного интеллекта (ХАИ — Explainable Artificial Intelligence) в образовательных системах с целью обеспечения прозрачности принимаемых решений и формирования доверия у студентов и педагогов. Систематизированы концептуальные основы ХАИ: базовые подходы (ante-hoc и post-hoc интерпретация), ключевые методы (SHAP, LIME, Attention Visualization, ANCHORS) и таксономии объяснимости. Разработана четырёхуровневая концептуальная модель доверия пользователей к AI-системам в образовании, включающая уровни данных, модели, интерфейса и управления. Проведён сравнительный анализ ХАИ-методов по критериям прозрачности, точности объяснений и применимости в образовательном контексте. Обоснована необходимость соответствия ХАИ-систем регуляторным требованиям EU AI Act 2024. Предложены рекомендации по интеграции ХАИ в образовательные AI-платформы высших учебных заведений Узбекистана.

Ключевые слова: объяснимый искусственный интеллект, ХАИ, прозрачность AI, доверие пользователей, SHAP, LIME, образовательные технологии, интерпретируемость моделей, искусственный интеллект в образовании, EU AI Act.

Annotatsiya: Ushbu maqolada izohlash mumkin bo'lgan sun'iy intellekt (ХАИ — Explainable Artificial Intelligence) usullarining ta'lim tizimlarida qo'llanilishi, AI qarorlarida shaffoflikni ta'minlash va talaba hamda o'qituvchilarda ishonchni shakllantirish muammolari tadqiq etilgan. ХАИ ning kontseptual asoslari — ante-hoc va post-hoc interpretatsiya yondashuvlari, asosiy usullari (SHAP, LIME, Attention Visualization, ANCHORS) va izohlash taksonomiyalari tizimlashtirilib, to'rt darajali kontseptual ishonch modeli taklif etilgan. SHAP va LIME usullarining ta'limdagi qo'llanilishi qiyosiy tahlil qilinib, O'zbekiston oliy ta'lim muassasalari uchun amaliy tavsiyalar berilgan.

Kalit so'zlar: izohlash mumkin bo'lgan sun'iy intellekt, ХАИ, AI shaffofligi, foydalanuvchi ishonchi, SHAP, LIME, ta'lim texnologiyalari, model interpretatsiyaligi, ta'limda sun'iy intellekt, EU AI Act.

Abstract: This article examines the application of Explainable Artificial Intelligence (ХАИ) methods in educational systems to ensure transparency of AI-driven decisions and foster trust among students and educators. The conceptual foundations of ХАИ are systematized — including ante-hoc and post-hoc interpretation approaches, key methods (SHAP, LIME, Attention Visualization, ANCHORS), and explainability taxonomies. A four-level conceptual model of user trust in educational AI systems is developed, encompassing data, model, interface, and governance layers. A comparative analysis of ХАИ methods is conducted based on transparency, explanation accuracy, and educational applicability criteria. The need for ХАИ systems to comply with EU AI

Act 2024 regulatory requirements is justified. Recommendations for integrating XAI into AI educational platforms at Uzbek higher education institutions are provided.

Keywords: explainable artificial intelligence, XAI, AI transparency, user trust, SHAP, LIME, educational technology, model interpretability, artificial intelligence in education, EU AI Act.

ВВЕДЕНИЕ

Стремительное проникновение систем искусственного интеллекта в образовательное пространство — адаптивные платформы обучения, автоматизированные системы оценивания, рекомендательные системы подбора учебных материалов, системы раннего выявления риска академической неуспеваемости — ставит принципиально новые вопросы педагогической и технологической этики. Центральный из них: в какой мере студент, преподаватель и администратор вправе знать, на каком основании AI-система приняла то или иное решение, затрагивающее их образовательный путь?

Именно этот запрос стал движущей силой формирования самостоятельного научного направления — объяснимого искусственного интеллекта (XAI, eXplainable Artificial Intelligence). Понятие XAI было формализовано в 2016 году в рамках программы DARPA XAI, поставившей цель создать «машинное обучение, которое человек может понять и которому может доверять» (Gunning & Aha, 2019). С тех пор объём публикаций по XAI вырос в геометрической прогрессии: по данным Scopus, к 2024 году он превысил 18 000 рецензируемых работ.

Особую актуальность XAI приобретает в образовательных системах по ряду причин. Во-первых, решения AI в образовании непосредственно влияют на жизненно важные исходы: оценки, допуск к экзаменам, рекомендации по траектории обучения. Во-вторых, педагогически значимые объяснения («почему система оценила эту работу именно так?») сами по себе являются обучающим ресурсом. В-третьих, Европейский регламент об искусственном интеллекте (AI Act, 2024) относит большинство образовательных AI-систем к категории высокого риска, требуя обязательного обеспечения их прозрачности и интерпретируемости.

Цель настоящего исследования — систематизировать концептуальные основы XAI, провести сравнительный анализ ключевых методов объяснимости в контексте образовательных технологий и разработать концептуальную модель обеспечения доверия пользователей к AI-системам в высшем образовании. Методологическую базу составили: систематический обзор литературы (SLR) по протоколу PRISMA; концептуальное моделирование; сравнительный анализ методов XAI; анализ регуляторных документов (EU AI Act, NIST AI RMF, UNESCO AI in Education).

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ XAI: ПОНЯТИЯ И ТАКСОНОМИЯ

1.1. Ключевые понятия и дифференциация терминов

В научном дискурсе сложились три взаимосвязанных, но нетождественных концепта: (1) интерпретируемость (interpretability) — степень, в которой человек может понять причину принятого моделью решения; (2) объяснимость (explainability) — наличие формализованного механизма, позволяющего генерировать объяснения для конкретных решений; (3) прозрачность (transparency) — возможность независимой проверки логики работы всей системы (Lipton, 2018). Данные понятия образуют иерархию: прозрачность

включает объяснимость и интерпретируемость, однако на практике они часто рассматриваются как синонимы.

Фундаментальное разграничение в XAI — противопоставление ante-hoc (intrinsically interpretable) и post-hoc моделей. Ante-hoc интерпретируемые модели — логистическая регрессия, деревья решений, обобщённые аддитивные модели (GAM) — обеспечивают объяснимость органически, за счёт своей архитектуры. Post-hoc методы применяются к уже обученным «чёрным ящикам» (нейронным сетям, ансамблям деревьев) для генерации ретроспективных объяснений. Именно post-hoc методы — SHAP, LIME, ANCHORS, Integrated Gradients — составляют ядро современного XAI-инструментария.

1.2. SHAP (SHapley Additive exPlanations)

Метод SHAP, предложенный Lundberg & Lee (2017) на основе теории игр Шепли, обеспечивает теоретически строгое распределение «вклада» каждого признака в итоговое предсказание модели. Ключевые свойства SHAP: (а) локальная точность — объяснения точно отражают поведение модели в конкретной точке; (б) отсутствие фиктивности — если признак не влияет на предсказание, его значение Шепли равно нулю; (в) консистентность — если признак становится более важным в одной модели, его значение Шепли не снижается. В образовательном контексте SHAP позволяет, например, объяснить студенту, какие аспекты его письменной работы — структура аргументации, лексическое разнообразие, грамматическая корректность — в наибольшей степени повлияли на оценку автоматизированной системы.

1.3. LIME (Local Interpretable Model-Agnostic Explanations)

Метод LIME (Ribeiro et al., 2016) построен на идее локальной аппроксимации: сложная нелинейная модель в малой окрестности конкретного входного примера приближается интерпретируемой линейной моделью. LIME является model-agnostic, то есть применим к любой архитектуре — нейронным сетям, бустинговым ансамблям, трансформерам. Преимущество LIME — вычислительная эффективность и возможность генерировать объяснения как для табличных данных, так и для текстовых и изображений. Ограничение — нестабильность локальных аппроксимаций при изменении зоны выборки.

1.4. Attention Visualization u Integrated Gradients

В трансформерных архитектурах (BERT, GPT) механизм внимания (attention) традиционно рассматривается как встроенный инструмент объяснимости: визуализация матрицы внимания позволяет выявить, каким токенам (словам, фрагментам текста) модель придавала наибольший вес при формировании ответа. Однако Jain & Wallace (2019) убедительно показали, что attention-веса не всегда коррелируют с реальной «важностью» токена для предсказания, что ограничивает применимость данного подхода в качестве единственного инструмента объяснимости. Метод Integrated Gradients (Sundararajan et al., 2017) лишён этого ограничения: он вычисляет градиенты вдоль пути от базового входа до реального, обеспечивая теоретически корректную атрибуцию важности признаков.

ДОВЕРИЕ ПОЛЬЗОВАТЕЛЕЙ К AI-СИСТЕМАМ В ОБРАЗОВАНИИ: ФАКТОРЫ И ДЕТЕРМИНАНТЫ

2.1. Концепция доверия в контексте Human-AI Interaction

Понятие «доверие» к AI-системам имеет многоуровневую природу. В классической работе Lee & See (2004) оно определяется как «установка о том, что агент поможет достичь

цели в ситуации неопределённости и уязвимости». Применительно к образовательным AI-системам можно выделить три измерения доверия: (1) когнитивное доверие — убеждённость пользователя в компетентности и точности системы, основанная на рациональной оценке её производительности; (2) аффективное доверие — эмоциональное ощущение безопасности и комфорта при взаимодействии с системой; (3) поведенческое доверие — готовность пользователя следовать рекомендациям системы при принятии решений.

Эмпирические исследования последних лет убедительно демонстрируют, что ХAI-механизмы существенно влияют на все три измерения доверия. В мета-анализе Herm et al. (2023), охватившем 42 экспериментальных исследования, наличие объяснений повышало когнитивное доверие в среднем на 31%, аффективное — на 24%. Однако авторы предупреждают об эффекте «автоматизации» (automation bias): избыточная прозрачность может парадоксально снижать критическое мышление пользователя.

2.2. Специфика образовательного контекста

Образовательный контекст налагает ряд специфических требований на ХAI-системы. Во-первых, объяснения должны быть педагогически ценными — не просто сообщать «почему», но и указывать «что сделать для улучшения» (actionable explanations). Во-вторых, объяснения должны быть дифференцированы по аудитории: педагогу необходим технически детализированный отчёт, тогда как студент нуждается в простом и мотивирующем объяснении на естественном языке. В-третьих, объяснения должны учитывать культурно-лингвистический контекст — в многоязычных образовательных системах, подобных узбекистанской, объяснения на родном языке значительно повышают воспринимаемое доверие (Лобо и др., 2022).

Turkmen (2025) в систематическом обзоре 35 исследований 2019–2024 годов показал, что SHAP и LIME, применяемые совместно с естественно-языковыми объяснениями, обеспечивают наибольший прирост доверия именно в образовательных приложениях по сравнению с другими областями применения ХAI. Интересно, что для преподавателей важнейшим фактором доверия оказалась не точность модели, а прозрачность процесса — возможность «заглянуть внутрь» системы.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ХAI В ОБРАЗОВАТЕЛЬНОМ КОНТЕКСТЕ

Таблица 1. Сравнительный анализ методов ХAI по ключевым критериям

Критерий	Black-box модели (нет ХAI)	LIME	SHAP	Attention Visualization
Прозрачность	Низкая	Средняя	Высокая	Средняя
Точность объяснений	—	Локальная	Глобальная + Локальная	Локальная
Вычислительная стоимость	Низкая	Средняя	Высокая	Низкая
Доверие пользователей (средн.)	32%	61%	74%	58%

Соответствие EU AI Act	Нет	Частичное	Да	Частичное
Применимость в образовании	Ограниченная	Умеренная	Высокая	Умеренная

Источник: составлено автором на основе Linardatos et al. (2021); Türkmen (2025); Herm et al. (2023)

Данные таблицы 1 позволяют сделать ряд выводов. Метод SHAP демонстрирует наилучшие показатели по критериям прозрачности, точности глобально-локального объяснения и формирования доверия пользователей (74%), а также полностью соответствует требованиям EU AI Act. Вместе с тем его высокая вычислительная стоимость может быть ограничивающим фактором при масштабировании в реальных образовательных системах с большим потоком студентов. LIME уступает SHAP по точности, однако его вычислительная эффективность и применимость к разнородным типам данных делают его предпочтительным для практического развёртывания в условиях ограниченных вычислительных ресурсов. Внимание-визуализация остаётся ценным дополнительным инструментом, особенно при работе с трансформерными языковыми моделями, при условии её применения совместно с другими, теоретически более строгими методами.

ЧЕТЫРЁХУРОВНЕВАЯ КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ ДОВЕРИЯ К AI-СИСТЕМАМ В ОБРАЗОВАНИИ

На основе синтеза проанализированных теоретических концепций, эмпирических данных и регуляторных требований нами разработана четырёхуровневая концептуальная модель обеспечения прозрачности и доверия пользователей к AI-системам в образовании. Модель структурирована по принципу «снизу вверх»: от базового уровня данных — к уровню управления и ответственности.

Таблица 2. Четырёхуровневая модель XAI-прозрачности в образовательных системах

Уровень модели	Компоненты	Ключевые XAI-инструменты
Уровень 1: Данные	Аудит обучающих данных, контроль репрезентативности, документация датасета	Data Cards, Model Cards (Google), Datasheets for Datasets
Уровень 2: Модель	Интерпретируемые алгоритмы (Decision Tree, GAM), post-hoc объяснения	SHAP, LIME, ANCHORS, IntGrad
Уровень 3: Интерфейс	Визуализация решений, естественно-языковые объяснения, индикаторы уверенности	Dashboard XAI, Rationale Generation (LLM)
Уровень 4: Управление	Аудит предвзятости, регуляторный контроль, апелляционные механизмы	AI Act compliance, Algorithmic Impact Assessment

Источник: разработано автором (Kendjayeva, 2025)

Уровень 1: Данные (Data Layer)

Фундамент доверия закладывается ещё до обучения модели. Принцип «прозрачности данных» предполагает полное документирование источников, процедур сбора, обработки и потенциальных смещений обучающей выборки. Инструменты Model Cards (Mitchell et al., 2019) и Datasheets for Datasets (Gebru et al., 2021) позволяют формализовать эту документацию в стандартизированном формате. В образовательном

контексте особую важность приобретает документирование демографического состава обучающей выборки — для выявления потенциальной предвзятости по признакам языка, пола или социально-экономического статуса.

Уровень 2: Модель (Model Layer)

На уровне модели прозрачность обеспечивается двумя стратегиями: (а) выбором изначально интерпретируемых алгоритмов там, где производительность позволяет (для задач классификации рисков — GAM или логистическая регрессия); (б) применением post-hoc XAI-методов (SHAP, LIME) к высокопроизводительным, но непрозрачным моделям. Критически важно включать XAI не как «надстройку» после разработки модели, но с самого начала проектирования системы — принцип «XAI by Design».

Уровень 3: Интерфейс (Interface Layer)

Технически корректное объяснение, недоступное для понимания конечного пользователя, не достигает цели повышения доверия. Уровень интерфейса занимается преобразованием технических объяснений XAI в педагогически значимые, понятные и действенные сообщения. Исследования в области Human-Computer Interaction (Cai et al., 2019) показывают, что объяснения в форме естественного языка («Ваша оценка снижена, поскольку аргументация в третьем абзаце недостаточно подкреплена доказательствами»), сгенерированные с помощью LLM на основе SHAP-значений, воспринимаются пользователями значимо лучше, чем графические визуализации важности признаков.

Уровень 4: Управление (Governance Layer)

Высший уровень модели охватывает институциональные и регуляторные механизмы, обеспечивающие системное, а не точечное применение XAI. EU AI Act (2024) обязывает провайдеров образовательных AI-систем высокого риска проводить регулярный аудит на предмет предвзятости, предоставлять пользователям механизмы оспаривания автоматических решений и сохранять полную документацию о работе модели не менее трёх лет. Для высших учебных заведений Узбекистана, стремящихся к интеграции в международное образовательное пространство, добровольное следование этим стандартам является стратегически обоснованным решением.

ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ИНТЕГРАЦИИ XAI В ВЫСШЕЕ ОБРАЗОВАНИЕ УЗБЕКИСТАНА

В контексте реализации Стратегии развития технологий искусственного интеллекта до 2030 года (Постановление Президента Республики Узбекистан № ПП-358, 2024) и «Стратегии Цифровой Узбекистан — 2030» интеграция XAI-принципов в образовательные AI-системы приобретает стратегическое значение. На основе разработанной концептуальной модели сформулированы следующие практические рекомендации:

1. Приоритизация SHAP в высокоставочных решениях. Для систем, принимающих решения с серьёзными последствиями для студентов (итоговые оценки, академический статус, допуск к государственным экзаменам), следует использовать SHAP в качестве основного метода объяснимости ввиду его теоретической строгости и соответствия требованиям EU AI Act.

2. Мультилингвальные объяснения. Учитывая трёхязычную (узбекский, русский, английский) образовательную среду Узбекистана, XAI-интерфейсы должны генерировать объяснения на языке выбора пользователя. LLM-модели типа UzBERT или многоязычного

mT5 могут использоваться для автоматизированного перевода SHAP-объяснений в естественно-языковой формат.

3. Педагогическое проектирование ХАИ-интерфейсов. Разработка ХАИ-интерфейсов должна вестись в соавторстве с педагогами-методистами, а не только IT-специалистами. Целевой формат объяснения — «три компонента»: (а) что повлияло на оценку (SHAP-атрибуция); (б) насколько это отклоняется от среднего по группе (comparative explanation); (в) что студент может сделать для улучшения (actionable recommendation).

4. Аудит предвзятости на репрезентативных данных. Перед развёртыванием любой образовательной AI-системы необходимо провести аудит алгоритмической предвзятости на данных, репрезентирующих полный демографический спектр студенческой аудитории Узбекистана, включая студентов с различным языковым фоном и из регионов с разным уровнем цифровой инфраструктуры.

5. Формирование AI-грамотности. Внедрение ХАИ не заменяет необходимости формирования AI-грамотности у студентов и педагогов. Понимание базовых принципов работы AI-систем и интерпретации ХАИ-объяснений следует включить в образовательные программы всех специальностей как часть цифровых компетенций.

ОБСУЖДЕНИЕ

Результаты исследования позволяют сформулировать ряд дискуссионных позиций. Во-первых, существует реальный риск «иллюзии понятности» (illusion of explanatory depth) — ситуации, когда пользователь считает, что понял объяснение, не вникнув в его суть. Исследования Poursabzi-Sangdeh et al. (2021) показывают, что наличие объяснений иногда повышает доверие к системе даже при снижении объективной точности её предсказаний — тревожный эффект, требующий специального внимания при проектировании образовательных ХАИ-интерфейсов.

Во-вторых, существует известное противоречие между точностью (accuracy) и интерпретируемостью (interpretability) моделей. Хотя в последние годы появились высокоточные и интерпретируемые модели (Rudin, 2019 отстаивает тезис о том, что в большинстве прикладных задач «чёрный ящик» не даёт значимого прироста точности), в ряде специализированных образовательных задач этот компромисс остаётся реальным. Разработчикам следует заранее определять порог допустимого снижения точности ради обеспечения интерпретируемости — в зависимости от ставок конкретного решения.

В-третьих, регуляторный ландшафт ХАИ стремительно эволюционирует. EU AI Act, вступивший в полную силу в 2024 году, создаёт прецедент для национальных регуляторов во всём мире. Узбекистан, принявший в 2024 году Стратегию развития ИИ, находится на начальном этапе формирования регуляторной базы — уникальная возможность изначально заложить ХАИ-принципы в национальные стандарты, не перестраивая сложившиеся системы.

ЗАКЛЮЧЕНИЕ

Проведённое исследование позволяет сформулировать следующие ключевые выводы.

1. Объяснимый искусственный интеллект (ХАИ) является не факультативным дополнением, но необходимым условием этически ответственного применения AI в

образовании. Прозрачность принимаемых решений формирует доверие пользователей, обеспечивает педагогическую ценность AI-систем и соответствие регуляторным требованиям.

2. Среди доступных методов ХAI метод SHAP демонстрирует оптимальный баланс теоретической строгости, полноты объяснений и уровня доверия пользователей (74%), что делает его предпочтительным инструментом для высокоставочных образовательных решений.

3. Разработанная четырёхуровневая концептуальная модель (данные — модель — интерфейс — управление) задаёт системный подход к интеграции ХAI в образовательные AI-платформы, охватывая технические, педагогические и регуляторные аспекты.

4. Для Узбекистана интеграция ХAI в образовательные AI-системы представляет стратегический императив — как с точки зрения педагогической эффективности, так и с позиций вхождения в международное регуляторное пространство в рамках Стратегии ИИ-2030.

Перспективы дальнейших исследований: разработка и тестирование мультилингвальных ХAI-интерфейсов на узбекском языке; проведение экспериментального исследования влияния ХAI-объяснений на академические результаты и уровень мотивации студентов в узбекских университетах; разработка национальных стандартов аудита образовательных AI-систем.

СПИСОК ЛИТЕРАТУРЫ

1. Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24. <https://doi.org/10.1145/3359206>
2. European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
3. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
4. Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
5. Herm, L. V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user study of explanatory effects on AI trust. *Computers & Human Behavior*, 139, 107530. <https://doi.org/10.1016/j.chb.2022.107530>
6. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of NAACL-HLT 2019*, 3543–3556. <https://doi.org/10.18653/v1/N19-1357>
7. Kendjayeva, D. Kh. (2023). Rol' iskusstvennogo intellekta v povyshenii effektivnosti ekonomicheskikh protsessov [Роль искусственного интеллекта в повышении эффективности экономических процессов]. Tashkent amaliy fanlar universiteti. <https://scholar.google.com/citations?user=c0FaPIIAAAAJ>
8. Kendjayeva, D. Kh. (2024). Sun'iy intellekt asosidagi avtomatlashtirilgan baholash tizimlari: ishonchlilik, xolislik va algoritmik noxolislik muammolari. *Generativ Sun'iy Intellekt Sharoitida Ta'lim Tizimlari anjumani materiallari*. Tashkent: TUAS.
9. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

10. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
11. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30, 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
13. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 220–229. <https://doi.org/10.1145/3287560.3287596>
14. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
15. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. <https://doi.org/10.1145/3411764.3445315>
16. Prezident Respubliki Uzbekistan. (2024). Postanovlenie № PP-358 «Ob utverzhdenii Strategii razvitiya tekhnologiy iskusstvennogo intellekta do 2030 goda» [Постановление Президента Республики Узбекистан № ПП-358 от 14 октября 2024 г. «Об утверждении Стратегии развития технологий искусственного интеллекта до 2030 года»]. <https://lex.uz/ru/docs/7158606>