https://doi.org/10.31891/2307-5732-2026-365-70
УДК 004.8

ЛІП'ЯНІНА-ГОНЧАРЕНКО ХРИСТИНА
Західноукраїнський національний університет
https://orcid.org/0000-0002-2441-6292
e-mail:kh.lipianina@wunu.edu.ua
КОМАР МИРОСЛАВ
Західноукраїнський національний університет
https://orcid.org/0000-0001-6541-0359
e-mail: mko@wunu.edu.ua
БИКОВИЙ ПАВЛО
Західноукраїнський національний університет
https://orcid.org/0000-0002-5705-5702
e-mail: pb@wunu.edu.ua
ЮРКІВ ХРИСТИНА
Західноукраїнський національний університет
https://orcid.org/0009-0007-4917-3251
e-mail: kh.yurkiv@wunu.edu.ua

ІНТЕГРАЦІЙНА РАМКА ОПЕРАЦІЙНОЇ ВЕРИФІКАЦІЇ ВІДПОВІДАЛЬНОГО ШІ НА ОСНОВІ МІЖНАРОДНИХ СТАНДАРТІВ

У статті обґрунтовано та сформовано інтеграційну рамку операційної верифікації відповідального штучного інтелекту, яка узгоджує міжнародні стандарти й рамкові документи ISO/IEC, IEEE, OECD, UNESCO, NIST та підходи ЄС у єдиній логіці перевірки відповідності AI/ML-систем. Запропонована рамка усуває розрив між декларативними принципами та практикою оцінювання, формалізуючи відтворюваний ланцюг доказовості «принцип – вимога – контроль/тест – доказовий артефакт – управлінське рішення –моніторинг». Визначено ядро характеристик довірчості (робастність і безпечність, кіберстійкість, приватність та управління даними, прозорість і пояснюваність, справедливість і недискримінація, підзвітність і трасованість) та показано комплементарні ролі інституцій: ISO/IEC – процеси управління ризиками й менеджменту, IEEE 7000 – інженерні механізми «by design», OECD/UNESCO – ціннісно-нормативний горизонт, NIST – операційна ризик-орієнтована структура, ЄС – регуляторні очікування для високоризикових застосувань. Обґрунтовано профілювання оцінювання за рівнем ризику та вимоги до стандартизованого пакета доказів (паспорти моделі й даних, протоколи випробувань, реєстр ризиків і план пом'якшення, журнали експлуатації, умови людського нагляду, план моніторингу). Практичним результатом є рекомендації щодо архітектури національної/міжвідомчої платформи тестування відповідальності ШІ, сумісної з європейською логікою ризик-орієнтованої відповідності та аудитопридатності.

Ключові слова: *відповідальний штучний інтелект; довірливий ШІ; стандарти ISO/IEC; IEEE 7000; аудитопридатність; прозорість і пояснюваність; справедливість і недискримінація.*

LIPIANINA-HONCHARENKO KHRYSTYNA, KOMAR MYROSLAV, BYKOVYY PAVLO, YURKIV KHRYSTYNA
West Ukrainian National University

INTEGRATION FRAMEWORK FOR OPERATIONAL VERIFICATION OF RESPONSIBLE AI BASED ON INTERNATIONAL STANDARDS

The paper substantiates and develops an integration framework for the operational verification of responsible artificial intelligence, aligning international standards and framework documents of ISO/IEC, IEEE, OECD, UNESCO, NIST, and European Union regulatory approaches within a unified logic for assessing compliance of AI/ML systems. The study addresses the methodological gap between high-level ethical principles and their practical operationalization in measurable, reproducible, and auditable evaluation procedures. The proposed framework introduces a structured evidence chain: "principle – requirement – control/test – evidentiary artifact – managerial decision – post-market monitoring." This logic ensures traceability, comparability, and auditability of assessment results across different types of AI systems, including traditional machine learning models, computer vision systems, natural language processing models, and generative AI. The core trustworthiness characteristics are systematized as robustness and safety, cyber resilience, privacy and data governance, transparency and explainability, fairness and non-discrimination, accountability, and traceability throughout the entire AI lifecycle. The complementary institutional roles are clarified: ISO/IEC establishes risk management and AI management system processes; the IEEE 7000 series operationalizes ethical and "by design" engineering requirements; OECD and UNESCO define value-based and human-rights-oriented principles; NIST provides a risk-oriented governance structure; and the European Union translates these principles into regulatory obligations, particularly for high-risk AI systems. The framework incorporates risk-based profiling of evaluation procedures and defines a standardized evidence package, including model and data documentation, testing protocols, risk registers and mitigation plans, operational logs, human oversight requirements, and continuous monitoring mechanisms. As a practical contribution, the study outlines the architecture and minimum functional components of a national/interagency responsible AI testing platform compatible with European risk-based compliance logic. The proposed approach transforms responsible AI from a declarative concept into an operational, verifiable, and auditable governance practice that strengthens regulatory alignment and public trust in AI-driven decision-making.

Keywords: *Responsible artificial intelligence; trustworthy AI; ISO/IEC standards; IEEE 7000; auditability; transparency and explainability; fairness and non-discrimination.*

Стаття надійшла до редакції / Received 11.02.2026
Прийнята до друку / Accepted 11.03.2026
Опубліковано / Published 28.05.2026

Актуальність проблеми

Швидка інтеграція моделей штучного інтелекту в державні послуги, медіа, освіту, охорону здоров'я та

бізнес в Україні посилює потребу формалізованої, стандартизованої та відтворюваної перевірки відповідальності AI/ML-систем. Ідеться не лише про досягнення високих показників точності, а про гарантовану відповідність критичним характеристикам довірчості: безпеці та робастності, приватності, недискримінації, пояснюваності, кіберстійкості та підзвітності протягом усього життєвого циклу системи – від проєктування і навчання до експлуатації та моніторингу [1–3,8]. Водночас, попри наявність розвиненого міжнародного стандартного та нормативно-методичного поля (ISO/IEC JTC 1/SC 42, серія IEEE 7000, OECD, UNESCO, NIST AI RMF), для організацій практичною проблемою залишається операціоналізація цих вимог у вигляді вимірюваних тестів, уніфікованих протоколів аудиту та доказових артефактів, придатних для різних типів моделей (табличні ML, CV, NLP, LLM/генеративні та мультимодальні) [1,2,4,5,7].

Додаткову актуальність проблемі надає європейський регуляторний контекст: розвиток механізмів відповідності та ризик-орієнтованих підходів у ЄС формує зростаючий попит на прозору перевірку ризиків і пакети доказів для організацій, інтегрованих у ринки та ланцюги постачання ЄС [2,3,6,8]. Однак на рівні практики в Україні спостерігається розрив між: (а) наявними міжнародними вимогами та рамками і (б) реальними процедурами перевірки, які часто мають фрагментарний характер (перевірка окремих метрик без єдиного “контуру доказовості”, без трасованості рішень та без формалізованого управління ризиками). Це унеможливлює порівнюваність результатів оцінювання між організаціями, послаблює аудитопридатність і створює ризики впровадження моделей без належної валідації у чутливих доменах [1–3,8].

Отже, ключова науково-прикладна проблема полягає у відсутності узгодженої інтеграційної рамки та інструментального контуру, що об’єднює міжнародні стандарти і фреймворки у єдину структуру *«принцип – вимога – контроль/тест – доказовий артефакт – управлінське рішення –моніторинг»* [1–5,7,8]. Саме така рамка є передумовою створення платформи стандартизованого тестування відповідальності ШІ та формування доказової бази відповідності, яка забезпечуватиме відтворюваність оцінювання, прозорість ризиків і підзвітність рішень у процесі впровадження AI/ML-систем [1–3,6,8].

Аналіз останніх досліджень і публікацій.

Сучасний корпус публікацій у сфері відповідального ШІ демонструє чіткий зсув від декларативних принципів до інженерної реалізації вимог через керовані ризиками процеси, метрики та контрольні процедури. У стандартизованому вимірі ISO/IEC TR 24028:2020 формує “каркас довірчості” (reliability/robustness/safety/security/privacy/transparency/fairness/accountability) як спільну основу для оцінювання систем [1], тоді як ISO/IEC 23894:2023 конкретизує ризик-менеджмент ШІ як інтегровану організаційну практику [3], а ISO/IEC 42001:2023 переводить ці вимоги в площину системи менеджменту ШІ (політики, процедури, PDCA-цикл, доказовість та простежуваність) [8]. Паралельно IEEE 7000-2021 пропонує процесну модель інтеграції етичних цінностей у розробку (value elicitation – traceability – design dispositions) [4], що є методологічно важливим для формування тестових критеріїв і артефактів, які можуть бути перевірені незалежно.

У наукових виданнях, індексованих у Scopus, домінує лінія досліджень, яка з’єднує принципи, регуляцію та інженерні вимоги. Так, Díaz-Rodríguez та співавт. систематизують “ланцюг відповідальності” від етичних принципів до технічних/організаційних вимог і регуляторних очікувань, підкреслюючи потребу в узгоджених таксономіях вимог і валідаційних практиках для відповідального ШІ [10]. Для галузевих контекстів, де ризики максимально “матеріальні”, Moreno-Sánchez та співавт. пропонують дизайн-фреймворк операціоналізації Trustworthy AI у медицині як набір вимог і компромісів (напр., між пояснюваністю та продуктивністю, між приватністю та корисністю даних), що прямо вказує на необхідність стандартизованих тестових наборів і контекстних профілів ризику [11]. Додатково Wirz та співавт. критично переосмислюють поняття Trustworthy AI, підкреслюючи, що довірчість має не лише технічний, а й суб’єктивно-контекстний вимір: рівень довіри користувача залежить від ситуації застосування, ціни помилки та часових обмежень. Це методологічно важливо для проєктування платформ тестування, які мають підтримувати не тільки кількісні метрики, а й сценарні та людиноцентричні протоколи оцінювання [14].

У межах українського наукового дискурсу (фахові видання) останні роки характеризуються посиленням уваги до правових і правозахисних аспектів застосування ШІ – насамперед у частині персональних даних, відповідальності та судової/адміністративної практики. Заярний і Деркаченко на прикладі ChatGPT розкривають проблеми обробки персональних даних у взаємодії з ШІ-сервісами та акцентують на ризиках неконтрольованого розкриття/передавання даних і необхідності проєктування продуктів із вбудованими механізмами захисту персональних даних, що особливо актуально для тестування LLM і генеративних моделей [15]. Посикалюк аналізує баланс інтересів при використанні ШІ й персональних даних у площині європейських підходів та перспектив України, що підсилює висновок про потребу інституційних механізмів оцінки (а не лише загальних рекомендацій) [16]. У суміжній площині Ковальчук досліджує роль ШІ в судовій інтерпретації права, де критичною стає відтворюваність і пояснюваність результатів та доказовість процедур прийняття рішень – тобто ті виміри, які стандарти ISO/IEC трактують як складники довірчості й підзвітності [17]. Нарешті, Теремецький і Ковальчук розглядають ШІ як чинник цифрової трансформації правосуддя, фактично підкреслюючи потребу в процедурному контролі та механізмах нагляду при впровадженні алгоритмічних систем у чутливих доменах [14], тоді як Берназюк фокусується на викликах прав людини та міжнародному/українському досвіді, що задає нормативну “рамку очікувань” до національних інструментів оцінювання [18]. Сукупно ці роботи демонструють: українські публікації добре фіксують правові ризики та контури відповідальності, але системно бракує єдиного вимірювального контуру (метрик, тестів, протоколів, доказових артефактів), який би безпосередньо інтегрував

міжнародні стандарти (ISO/IEC, IEEE, NIST) у практику верифікації моделей усіх типів – від табличних ML до LLM і мультимодальних систем.

Критично важливо, що більшість сучасних рамок є модель-незалежними (застосовні до будь-яких ML/AI), але їх операціоналізація неминуче модель-специфічна. Для класичних моделей (LR/RF/XGBoost) основою тестування стають стабільність, узагальнення, групові метрики справедливості, аналіз чутливості та аудит даних; для комп'ютерного зору (CNN/ViT) – робастність до збурень/зсувів домену та помилок у розмітці; для NLP/BERT-подібних – зсуви мови/тематики, токсичність і упередження; для LLM/генеративних і мультимодальних – ризики галюцинацій, ін'єкцій у промпти, витоків даних, відтворення шкідливого контенту, а також вимоги прозорості щодо даних/авторського права, які підсилюються регуляторикою ЄС [12–13]. Тому актуальна тенденція досліджень полягає в переході до управління тестуванням: платформа оцінювання має поєднувати стандарти процесу (ISO/IEC 42001, ISO/IEC 23894) [3,8], контрольні списки вимог (ALTAI) [9], та ризик-орієнтовані функції керування (NIST AI RMF: Govern–Map–Measure–Manage) [2], формуючи відтворювані доказові матеріали відповідальності.

Метою статті є обґрунтувати та сформувати інтеграційну рамку оцінювання відповідального ШІ, яка узгоджує міжнародні стандарти й фреймворки (ISO/IEC, IEEE, OECD, ЄС, NIST, UNESCO) в єдиній логіці операційної верифікації: від принципів і вимог – до контрольних процедур, доказових артефактів, управлінських рішень і моніторингу, з урахуванням потреб гармонізації України з підходами ЄС.

Завдання дослідження;

1. Систематизувати міжнародні стандарти та рамки відповідального ШІ, виділивши узгоджене ядро характеристик *trustworthiness* (безпека/робастність, кіберстійкість, приватність, прозорість/пояснюваність, справедливість/недискримінація, підзвітність/трасованість) і роль кожної інституції (ISO/IEC, IEEE, OECD/UNESCO, ЄС, NIST) у багаторівневій архітектурі.

2. Розробити структуру інтеграційної рамки та формалізувати ланцюг доказовості, включно з профілями оцінювання за ризиком і вимогами до пакету доказів.

3. Обґрунтувати архітектуру та мінімальний функціональний набір національної/міжвідомчої платформи тестування відповідальності ШІ, сумісної з європейською логікою ризик-орієнтованої відповідності.

Наукова новизна. Запропоновано інтеграційну рамку операційної верифікації відповідального ШІ, що узгоджує міжнародні стандарти й фреймворки в єдиній логіці для підвищення аудитопридатності оцінювання.

Основні результати

Міжнародна організація зі стандартизації (ISO) спільно з Міжнародною електротехнічною комісією (IEC) створили підкомітет JTC 1/SC 42, який займається стандартами в галузі ШІ. Цей підкомітет розробив низку документів, що визначають вимоги до відповідального ШІ протягом усього життєвого циклу системи – від проєктування до використання і моніторингу. Ключовим поняттям є «довірливий ШІ», тобто такий, що заслуговує довіри користувачів і регуляторів.

У технічному звіті ISO/IEC TR 24028:2020 наведено огляд критеріїв довірчості систем ШІ, зокрема: надійність та відтворюваність результатів, стійкість до збоїв, безпечність, захищеність від кібератак, приватність даних, прозорість і пояснюваність, справедливість (відсутність упередженої дискримінації) та підзвітність (наявність відповідальних осіб і документованих доказів рішень) [1]. Цей документ не є формальним стандартом для сертифікації, але встановлює спільний словник і концепції, за якими можна оцінювати, чи відповідає система ШІ критеріям довіри [1]. На основі цих критеріїв у подальшому розробляються методики управління ШІ.

ISO також випустила стандарти, що регулюють ризики та процеси управління ШІ. Зокрема, ISO/IEC 23894:2023 визначає підходи до управління ризиками ШІ, аби забезпечити безпечність, надійність і відповідність регуляторним вимогам систем ШІ [19]. Для належного корпоративного управління ШІ існує стандарт ISO/IEC 38507:2022, що описує принципи управління і контролю використання ШІ в організаціях, аби забезпечити етичне впровадження і належний нагляд за алгоритмами [19]. Комплексні вимоги до управлінської системи в галузі ШІ встановлює ISO/IEC 42001:2023 – стандарт на систему менеджменту ШІ, аналогічний до ISO 9001, але з фокусом на відповідальне управління життєвим циклом рішень ШІ [19].

Окремі технічні звіти ISO/IEC спрямовані на специфічні аспекти відповідального ШІ. Так, ISO/IEC TR 24027:2021 присвячено виявленню та зменшенню упередженості в алгоритмах і рішеннях ШІ, щоб запобігти дискримінації [20]. Документ ISO/IEC TR 24368:2022 містить огляд етичних та соціальних питань у сфері ШІ як відправну точку для політик, що мають на меті захистити права споживачів і суспільство [19]. Загалом, стандарти ISO/IEC формують комплексну базу для інтеграції принципів етики, прозорості та безпеки у створення й експлуатацію моделей штучного інтелекту/машинного навчання (ШІ/МН).

Інститут інженерів з електротехніки та електроніки (IEEE) ініціював програми для забезпечення етичності та відповідальності ШІ. У межах глобальної ініціативи IEEE з етики автономних та інтелектуальних систем було підготовлено рекомендації «Ethically Aligned Design» (2019), що окреслили високорівневі принципи (пріоритет прав людини, прозорість, підзвітність, приватність, добробут суспільства тощо). На їх основі IEEE розробляє серію спеціалізованих стандартів IEEE 7000, кожен з яких фокусується на окремому аспекті відповідального ШІ.

Першим у цій серії став стандарт IEEE 7000-2021, який визначає модель процесу врахування етичних чинників при проєктуванні систем [20]. Він надає інженерам методику інтеграції етичного аналізу в життєвий цикл розробки технологій, щоби вже на етапі проєктування виявляти та вирішувати потенційні етичні дилеми.

Далі серія конкретизує вимоги за ключовими доменами: IEEE 7001-2021 визначає вимоги до прозорості автономних систем, встановлюючи вимірювані рівні прозорості [20]; IEEE 7002-2022 описує процес забезпечення приватності, вбудованої у проєктування («приватність на етапі проєктування»), а IEEE 7003-2022 – підходи до запобігання алгоритмічній упередженості та забезпечення справедливості [20]. Також виділяються стандарти щодо спеціальних категорій даних: IEEE 7004-2022 (дані дітей/учнів) та IEEE 7005-2021 (дані працівників) [20].

Окремо заслуговує уваги стандарт IEEE 7008-2022, який встановлює вимоги до довіри до систем ШІ, фактично визначаючи критерії довірчості. Сукупно ці документи формують прикладний інструментарій для інженерів і розробників: дотримання стандартів IEEE допомагає компаніям впроваджувати етичний ШІ – прозорий, безпечний, неупереджений – і зміцнювати довіру користувачів та суспільства до рішень ШІ [20].

Одним із перших міжурядових підходів до відповідального ШІ стали Принципи OECP (2019) – рекомендації для урядів і учасників екосистеми ШІ, спрямовані на забезпечення інноваційності та водночас довіри до ШІ з повагою до прав людини і демократичних цінностей [5]. Принципи містять п’ять базових ціннісних положень:

– інклюзивне зростання, сталий розвиток та добробут;

– повага прав людини та справедливість (включно з людським наглядом);

– прозорість та підзвітність (пояснюваність, можливість оскарження);

– надійність, безпека та стійкість (включно з механізмами зупинки/корекції);

– підзвітність (простежуваність, документування, аудитопридатність).

Європейський Союз у 2019 році через незалежну групу HLEG опублікував Ethics Guidelines for Trustworthy AI, де визначено, що надійний ШІ має бути законним, етичним і технічно надійним [6], а також сформульовано 7 вимог: людський нагляд; технічна надійність і безпека; приватність та управління даними; прозорість; різноманіття/недискримінація/справедливість; суспільне та екологічне благополуччя; підзвітність. Для практичної самооцінки запропоновано інструмент ALTAI [6]. Паралельно ЄС рухається до нормативного регулювання через Акт про штучний інтелект (AI Act) із ризик-орієнтованою класифікацією систем та підвищеними вимогами для високоризикових застосувань [6, 20].

Національний інститут стандартів і технологій США (NIST) запропонував добровільну Рамку управління ризиками ШІ (AI RMF 1.0, 2023), що структурує управління ризиками через чотири функції: врядування, визначення контексту, вимірювання, управління [2]. Центральним елементом є довірчість: валідність/надійність, безпечність, кіберстійкість, підзвітність/прозорість, пояснюваність, приватність і справедливість [2]. Також NIST публікує спеціалізовані документи щодо упередженості та пояснюваності, які доповнюють загальну рамку [20].

UNESCO у 2021 році ухвалила Recommendation on the Ethics of Artificial Intelligence – глобальний документ для країн-членів, що задає етичний каркас на основі прав людини і людської гідності та формулює принципи «не нашкодь», безпеки, приватності, прозорості, справедливості, підзвітності, інклюзивності й екологічної стійкості. Рекомендація також містить політичні дії для урядів: механізми оцінки впливу, реєстри алгоритмів, розвиток освіти тощо [7].

Серед інших ініціатив варто згадати роботу Ради Європи, яка готує першу міжнародно-правову конвенцію про ШІ із фокусом на права людини, демократію і верховенство права, а також форуми WEF та GPAI, що публікують рекомендації й інструментарії для відповідального впровадження ШІ. Хоча ці ініціативи (Таблиця 1) не завжди набувають форми стандартів, вони доповнюють загальну картину, формуючи міжнародний консенсус щодо розвитку ШІ у спосіб, сумісний із етичними цінностями, безпекою і довірою.

Показані в Таблиці 1 стандарти та рамки демонструють, що міжнародна спільнота фактично сформувала багаторівневу архітектуру відповідального ШІ/МН, у якій різні інституції виконують комплементарні ролі. ISO/IEC кодифікує термінологію, життєвий цикл і процеси управління ризиками та системи менеджменту, що задають «технічний каркас» керування ШІ [1,19]. IEEE конкретизує інженерні механізми реалізації етики, прозорості, приватності та недискримінації через стандарти «вбудованості у проєктування» [20]. OECD і UNESCO встановлюють ціннісно-нормативний горизонт (права людини, «не нашкодь», підзвітність, інклюзія, екологічна стійкість), який підтримується на міжурядовому рівні [5,7]. ЄС, поєднуючи добровільні етичні настанови з обов’язковими вимогами регулювання (Акт про штучний інтелект), переводить цю систему координат у площину юридично значущих обов’язків для високоризикових застосувань [6]. NIST, у свою чергу, пропонує операційний, ризик-орієнтований підхід до управління довірчістю, сумісний із міжнародними стандартами [2].

Попри різний нормативний статус (стандарти, технічні звіти, рекомендації та рамкові настанови), спільна мета зазначених документів є концептуально узгодженою: забезпечити розвиток і впровадження ШІ у спосіб, що одночасно підтримує інновації та мінімізує соціально неприйнятні ризики – через прозорість, недискримінацію, безпеку, приватність, людський нагляд і підзвітність [1,2,5,7]. У цьому контексті критичною категорією стає довіра: прийнятність ШІ/МН для суспільства та регуляторів прямо залежить від наявності не лише декларативних принципів, а й доказової демонстрації того, що система відповідає визначеним вимогам протягом усього життєвого циклу [1,2,6,19].

Водночас саме тут проявляється методологічний розрив між «переліком принципів» і «технікою перевірки»: на практиці організації часто мають фрагментарні інструменти (окремі метрики, політики чи контрольні переліки), які не формують відтворюваного контуру оцінювання і не забезпечують аудитопридатної доказовості. Наукова новизна запропонованого підходу (рисунок 1) полягає в обґрунтуванні інтеграційної рамки,

яка синхронізує міжнародні стандарти й рамки в єдиній логіці операційної верифікації: від ціннісних принципів – до формальних вимог, тестів, артефактів і управлінських рішень [1,2,5,7,19,20].

Таблиця 1

Порівняння основних стандартів і фреймворків

Стандарт / рамка (рік) – тип – організація	**Ключові принципи / вимоги**	**Сфера застосування**
ISO/IEC TR 24028:2020 «Довірчість ШІ» – технічний звіт – ISO/IEC (JTC 1/SC 42)	Надійність і відтворюваність; робастність і безпечність; кіберзахищеність; приватність; прозорість і пояснюваність; справедливість (недискримінація); підзвітність і простежуваність [1]	Узагальнені критерії довірчості систем ШІ; спільна термінологічна база для оцінювання відповідального ШІ [1]
ISO/IEC 23894:2023 «Управління ризиками ШІ» – міжнародний стандарт – ISO/IEC (JTC 1/SC 42)	Ризик-орієнтоване управління ШІ: ідентифікація, аналіз, оцінювання, оброблення та моніторинг ризиків протягом життєвого циклу [19]	Організації-розробники/інтегратори/експлуатанти ШІ у різних сферах; формування відтворюваних процедур управління ризиками [19]
ISO/IEC 38507:2022 «Врядування ШІ в організаціях» – міжнародний стандарт – ISO/IEC (JTC 1/SC 40)	Розподіл відповідальності; наглядові структури; політики та контроль використання ШІ; етичність і підзвітність у корпоративному врядуванні [19]	Організації, що впроваджують/використовують ШІ (держсектор, бізнес тощо); інтеграція ШІ у систему управління організацією [19]
ISO/IEC 42001:2023 «Система менеджменту ШІ» – міжнародний стандарт – ISO/IEC (JTC 1/SC 42)	Вимоги до системи менеджменту ШІ (аналогічно логіці ISO 9001): політики, процедури, цикл PDCA, доказовість, простежуваність, постійне вдосконалення[19]	Побудова керованої та аудитопридатної системи управління життєвим циклом рішень ШІ в організаціях [19]
IEEE 7000-2021 «Етичні чинники під час проєктування систем» – стандарт – IEEE (SA)	Процес урахування етичних чинників у проєктуванні: виявлення етичних ризиків, узгодження цінностей, простежуваність рішень у дизайні [20]	Загальні ІТ/ШІ-системи; методична основа для інтеграції етики у розробку («етика на етапі проєктування») [20]
IEEE 7001-2021 «Прозорість автономних систем» – стандарт – IEEE (SA)	Вимірювані рівні прозорості; вимоги до пояснюваності та розкриття інформації користувачу/стейкхолдерам [20]	Автономні та інтелектуальні системи; підвищення прозорості функціонування і рішень [20]
IEEE 7002-2022 / IEEE 7003-2022 «Приватність даних / Алгоритмічна упередженість» – стандарти – IEEE (SA)	7002: «приватність на етапі проєктування»; 7003: запобігання упередженості, справедливість, недискримінація, вимоги до даних і процедур [20]	Системи ШІ, що працюють із персональними даними або впливають на людей; інструментарій для розробників і аудиторів [20]
Принципи OECP щодо ШІ (2019) – міжнародна рекомендація – OECD	Цінності: інклюзивність і добробут; права людини та справедливість; прозорість і можливість оскарження; надійність/безпека/стійкість; підзвітність і аудитопридатність [5]	Орієнтир для національних стратегій і практик у всіх секторах; рамка для урядів і учасників екосистеми ШІ [5]
Ethics Guidelines for Trustworthy AI (2019) – рекомендації – Європейська комісія (HLEG)	7 вимог: людський нагляд; технічна надійність і безпека; приватність і управління даними; прозорість; різноманіття/недискримінація/справедливість; суспільне та екологічне благо; підзвітність (ALTAI як інструмент самооцінювання) [6]	Добровільне застосування для організацій; практична рамка для самооцінювання та підготовки до регуляторних вимог ЄС [6]
NIST AI RMF 1.0 (2023) – рамкова настанова – NIST (США)	Функції: врядування, визначення контексту, вимірювання, управління; характеристики довірчості: валідність/надійність, безпечність, кіберстійкість, прозорість/підзвітність, пояснюваність, приватність, справедливість [2]	Добровільна рамка для організацій у різних сферах; побудова процесів керування ризиками та довірчістю ШІ [2]
UNESCO Recommendation on the Ethics of AI (2021) – міжнародна рекомендація – UNESCO (ООН)	Пріоритет прав людини; «не нашкодь»; пропорційність; безпека; приватність; прозорість; справедливість; підзвітність; інклюзивність; екологічна стійкість; політичні дії (оцінка впливу, реєстри, освіта)	Орієнтир для державної політики та етичного врядування ШІ на національному рівні; застосування у публічних і соціально чутливих сферах [7]

Запропонована рамка має чотирирівневу структуру та одночасно охоплює: (1) ціннісно-нормативний рівень прав людини, недискримінації, прозорості, безпеки, приватності та підзвітності, сформований у документах OECD і UNESCO [5,7]; (2) процесний рівень управління ризиками та менеджменту ШІ протягом життєвого циклу (ISO/IEC 23894:2023; ISO/IEC 42001:2023) [19]; (3) вимірювальний рівень довірчості як об'єкта оцінювання (кіберзахищеність, приватність, прозорість і пояснюваність, справедливість, підзвітність і простежуваність, робастність і безпечність), визначений у ISO/IEC TR 24028:2020 і узгоджуваний із Рамкою управління ризиками ШІ NIST [1,2]; (4) інженерний рівень реалізації вимог через підходи «вбудовані на етапі проєктування» (етика на етапі проєктування, приватність на етапі проєктування, запобігання упередженості, прозорість) на основі серії IEEE 7000 [20]. Така композиція формує зв'язок між принципами та контрольними процедурами, забезпечуючи порівнюваність результатів між організаціями і предметними областями.

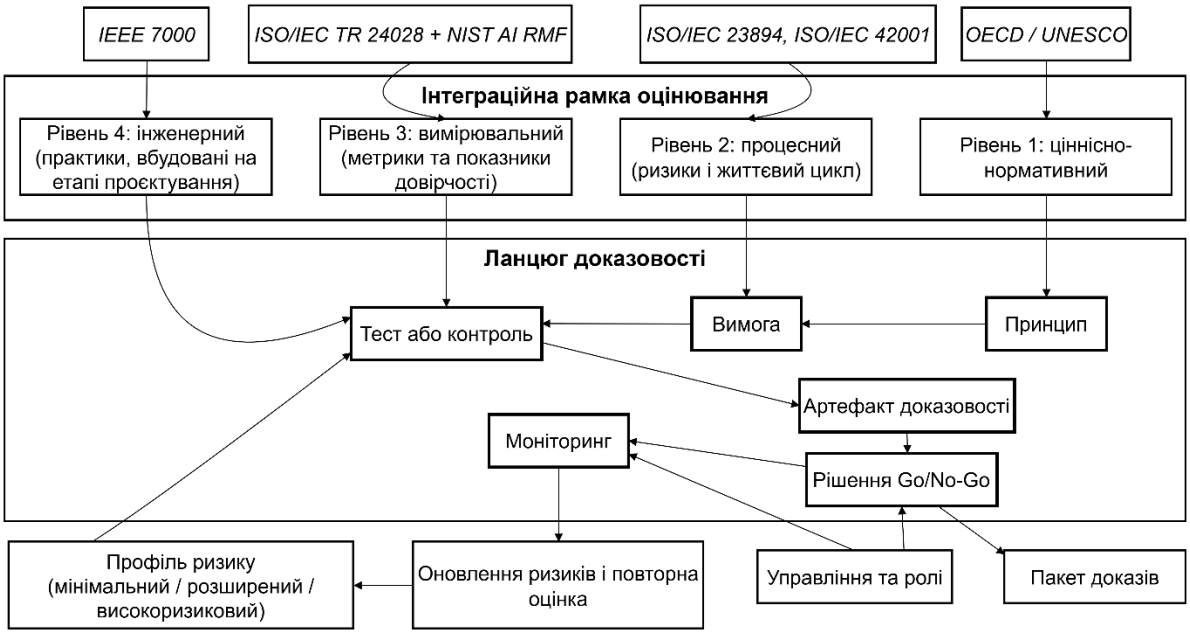

Рисунок 1 – Інтеграційна рамка оцінювання відповідального ШІ

Концептуально рамка вводить ланцюг доказовості: принцип – вимога – контроль/тест – доказовий артефакт – управлінське рішення – моніторинг. Її прикладна цінність полягає в тому, що кожна характеристика довірчості отримує не лише описову інтерпретацію, а й формалізований інструментарій перевірки та пакет доказів. Так, для справедливості визначальними стають протоколи аудиту упередженості та групові метрики справедливості; для прозорості – результати процедур пояснюваності та специфікації інформування користувача; для безпеки – звіти щодо стійкості до атак і зловживань; для приватності – документи управління даними та результати оцінки ризику витоку [1,2,5,7,19,20]. Таким чином, «довірчість» переходить із концептуального рівня у площину перевірюваних і трасованих результатів.

Процесна частина рамки операціоналізується через ризик-орієнтовану логіку «Врядування – Контекстуалізація – Вимірювання – Управління» (підхід NIST) у поєднанні з ISO-підходами до управління ризиками та системного менеджменту [2,19]. На вході формується профіль застосування (контекст, зацікавлені сторони, критичність рішення, припустимий рівень ризику), після чого визначається профіль оцінювання (мінімальний/розширений/високоризиковий) із наперед заданим обсягом тестів і вимог до документування. На виході рамка вимагає стандартизованого пакета доказів: паспорти моделі й даних, протоколи тестів і журнали виконання, реєстр ризиків і план їх пом'якшення, рішення про допуск/недопуск до впровадження, умови людського нагляду та план післявпроваджувального моніторингу [1,2,19]. Принципово важливо, що оцінюється не лише модель, а весь ланцюг «дані – навчання – інтеграція – експлуатація – моніторинг», без чого неможлива реальна підзвітність [1,2,5].

У контексті євроінтеграційного вектора України інтеграційна рамка виконує додаткову функцію – гармонізує доказовість із очікуваннями регуляторного середовища ЄС (у частині процесної зрілості, протоколів, артефактів та незалежного аудиту), що є критичним для високоризикових застосувань [1,2,6,7,19]. На цій основі обґрунтовується створення в Україні національної/міжвідомчої платформи тестування відповідальності ШІ як інфраструктурного елемента екосистеми відповідального ШІ, яка поєднує методики оцінювання довірчості [1,2], процеси управління ризиками й системи менеджменту [19], корпоративне врядування та розподіл відповідальності [19], інженерні стандарти IEEE 7000 [20] і міжурядові ціннісні рамки [5,7], будучи сумісною з логікою європейського регулювання [6].

Практичне проєктування такої платформи передбачає, по-перше, нормативно-методичну узгодженість через єдину матрицю критеріїв і показників (оціночну матрицю відповідального ШІ), що прямо співвідноситься з характеристиками довірчості [1,2], процесами ISO-управління ризиками та системами менеджменту [19] і вимогами IEEE 7000 щодо етики/приватності/запобігання упередженості/прозорості, інтегрованих у проєктування [20]. По-друге, необхідною є ризик-орієнтована класифікація як вхідний етап оцінювання, що формалізує контекст застосування, групи впливу, критичність рішення та допустимість ризику, відповідно до логіки NIST та ISO-підходів [2,19]. По-третє, платформа має забезпечувати оцінювання життєвого циклу, запроваджуючи обов'язкові артефакти доказовості (паспорт моделі, паспорт даних, протокол навчання, реєстр змін, журнали експлуатації, план моніторингу), що підсилює трасованість і відтворюваність [1,2,5,19].

По-четверте, на рівні функціоналу платформа повинна підтримувати мінімальний обов'язковий набір тестів відповідальності: справедливість/недискримінація [5,7,20], робастність і безпечність [1,2], кіберстійкість і захист [1,2], приватність і управління даними [1,6,21], прозорість/пояснюваність і інформування користувача [5,6,20]. Для генеративних моделей доцільним є окремий профіль тестування (відтворюваність контенту, підтвердження походження, ризики галюцинацій і сценарії зловживань), узгоджений із сучасною практикою

управління ризиками [2]. По-п’яте, результатом має бути не лише рейтинг, а стандартизована доказова папка (пакет доказів) з протоколами, журналами, межами застосовності, планом оброблення ризиків, рішенням про допуск/недопуск і вимогами до людського нагляду, що забезпечує аудитопридатність та зовнішню верифікацію [1,2,5,19]. По-шосте, платформа має технологічно фіксувати врядування і ролі (розробник, інтегратор, експлуатант, незалежний оцінювач, уповноважена посадова особа), підтримуючи незмінні журнали рішень та механізми оскарження/розгляду інцидентів відповідно до принципів підзвітності [5,7,19].

По-сьоме, післявпроваджувальний моніторинг має бути обов’язковим модулем: безперервний контроль показників якості й ризику, управління інцидентами, процедури обмеження функцій або відкликання версій у разі дрейфу даних чи появи нових загроз [1,2,19]. По-восьме, для масштабування необхідна інтеграція з держсектором, освітою та закупівлями, тобто включення вимоги проходження тестування та надання пакета доказів у технічні умови пілотів, грантових програм і державних закупівель у чутливих сферах. По-дев’яте, впровадження доцільно здійснювати поетапно через мінімально життєздатний прототип – розширення (генеративні моделі/високочутливі домени) – інституціоналізацію із залученням незалежних лабораторій, що одночасно знижує бар’єри для інновацій і закріплює культуру доказовості та ризик-орієнтованого управління [2,5,19].

Отже, запропонований підхід забезпечує операційну новизну: (1) вводить інтеграційну рамку, яка формалізує перехід від принципів до контрольних процедур, (2) визначає відтворюваний ланцюг доказовості для ключових характеристик довірчості, і (3) обґрунтовує архітектуру національної платформи тестування як інструмента гармонізації українських практик із міжнародними стандартами та європейською логікою ризик-орієнтованої відповідності [1,2,5-7,19,20]. Саме поєднання формалізованого тестового контуру з системою підзвітності та доказовості переводить відповідальний ШІ із декларативної площини в аудитопридатну інженерно-управлінську практику, зменшуючи регуляторну невизначеність та підвищуючи суспільну довіру до рішень ШІ/МН.

Висновки

У межах виконання поставлених завдань систематизовано міжнародні стандарти та рамкові документи відповідального ШІ (ISO/IEC, IEEE, OECD, ЄС, NIST, UNESCO) і встановлено їх концептуальну узгодженість навколо ядра характеристик довірчості: надійність і відтворюваність, робастність і безпечність, кіберстійкість, приватність та управління даними, прозорість і пояснюваність, справедливість і недискримінація, підзвітність і простежуваність. Показано взаємодоповнюваність інституційних ролей: ISO/IEC формує процесно-управлінський каркас життєвого циклу та управління ризиками, IEEE конкретизує інженерні практики «на етапі проєктування», OECD і UNESCO задають ціннісно-нормативний горизонт, ЄС переводить принципи у площину юридично значущих вимог, тоді як NIST пропонує операційний ризик-орієнтований підхід до управління довірчістю.

На основі виявлених відповідностей обґрунтовано інтеграційну рамку оцінювання відповідального штучного інтелекту, що усуває методологічний розрив між декларативними принципами та технікою перевірки, переводячи довірчість у площину перевірюваних і відтворюваних результатів. Рамка формалізує ланцюг доказовості «принцип – вимога – контроль/випробування – доказовий артефакт – управлінське рішення – моніторинг», передбачає профілювання оцінювання за рівнем ризику та визначає стандартизований склад пакета доказів (паспорти моделі й даних, протоколи випробувань і журнали виконання, реєстр ризиків і план їх пом’якшення, рішення «дозволити/заборонити впровадження», умови людського нагляду та план післявпроваджувального моніторингу).

Практичним наслідком є обґрунтування архітектури національної/міжвідомчої платформи тестування відповідальності ШІ, сумісної з європейською логікою ризик-орієнтованої відповідності та вимогами аудитопридатності. Визначено ключові модулі платформи: єдина матриця критеріїв і показників (оціночна картка відповідального ШІ), класифікація ризику, оцінювання повного життєвого циклу, мінімальний набір випробувань (справедливість/недискримінація; робастність/безпечність; кіберзахищеність; приватність; прозорість/пояснюваність), формалізоване врядування та розподіл ролей і відповідальності, незмінні журнали рішень, управління інцидентами та обов’язковий післявпроваджувальний моніторинг. Таким чином, запропонований підхід забезпечує операційну інтеграцію міжнародних норм у відтворюваний контур верифікації відповідального ШІ та підсилює доказовість і довіру до рішень ШІ у чутливих сферах.

Література

1. ISO/IEC TR 24028:2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. — Geneva : International Organization for Standardization, 2020.

2. NIST AI 100-1:2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). — Gaithersburg, MD: National Institute of Standards and Technology, 2023. — 48 p. — URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

3. ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. — Geneva : International Organization for Standardization, 2023.

4. EEE Std 7000-2021. IEEE Standard Model Process for Addressing Ethical Concerns During System Design. — New York : IEEE Standards Association, 2021.

5. OECD. Recommendation of the Council on Artificial Intelligence. — Paris : Organisation for Economic Co-operation and Development, 2019 (upd. 2024). — URL: https://oecd.ai/en/ai-principles

6. European Commission. Ethics Guidelines for Trustworthy AI. — Brussels : High-Level Expert Group on Artificial Intelligence, 2019. — URL: https://digital-strategy.ec.europa.eu

7. UNESCO. Recommendation on the Ethics of Artificial Intelligence. — Paris : United Nations Educational, Scientific and Cultural Organization, 2021. — URL: https://unesdoc.unesco.org

8. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. — Geneva : International Organization for Standardization, 2023.

9. European Commission. Assessment List for Trustworthy Artificial Intelligence (ALTAI). — Brussels : High-Level Expert Group on Artificial Intelligence, 2020. — URL: https://digital-strategy.ec.europa.eu

10. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation [Електронний ресурс] / Natalia Díaz-Rodríguez [та ін.] // Information Fusion. – 2023. – С. 101896. – Режим доступу: https://doi.org/10.1016/j.inffus.2023.101896

11. Moreno-SÃ, P. A., Del Ser, J., Van Gils, M., & Hernesniemi, J. (2025). A design framework for operationalizing trustworthy artificial intelligence in healthcare: Requirements, tradeoffs and challenges for its clinical adoption. Information Fusion, 103812.Regulation (EU) 2024/1689. Artificial Intelligence Act.

12. European Commission (2024). AI Act enters into force (повідомлення ЄК).

13. (Re)Conceptualizing trustworthy AI: A foundation for change [Електронний ресурс] / Christopher D. Wirz [та ін.] // Artificial Intelligence. – 2025. – Т. 342. – С. 104309. – Режим доступу: https://doi.org/10.1016/j.artint.2025.104309

14. Заярний, О. А., & Деркаченко, Ю. В. (2023). Деякі особливості обробки персональних даних при використанні чат-ботів зі штучним інтелектом на прикладі ChatGPT. Юридичний бюлетень, 55-62.

15. Посикалюк О. Штучний інтелект і персональні дані: до питання про пошук балансу інтересів [Електронний ресурс] / Олег Посикалюк // Право України. – 2025. – № 2025/07. – С. 122. – Режим доступу: https://doi.org/10.33498/louu-2025-07-122

16. Кобко-Одарій В. С. РОЛЬ ШТУЧНОГО ІНТЕЛЕКТУ В СУДОВІЙ ІНТЕРПРЕТАЦІЇ ПРАВА [Електронний ресурс] / В. С. Кобко-Одарій // Kyiv Law Journal. – 2023. – № 3. – С. 7–13. – Режим доступу: https://doi.org/10.32782/klj/2023.3.1

17. Teremetskyi V. Artificial Intelligence as a Factor in the Digital Transformation of the Justice System [Електронний ресурс] / V.I. Teremetskyi, O.Ya. Kovalchuk // Форум Права. – 2024. – Т. 78, № 1. – С. 106–115. – Режим доступу: https://doi.org/10.5281/zenodo.10870779

18. Habib P. E. M. AI Standards by ISO/IEC, AAMI & IEEE: Ensuring Ethical & Trustworthy AI [Електронний ресурс] / Prof Engr Murad Habib // LinkedIn: Log In or Sign Up. – Режим доступу: http://www.linkedin.com/pulse/ai-standards-isoiec-aami-ieee-ensuring-ethical-habib-xbqif#:~:text=4,of%20false%

19. Lopez O. AI ISO/IEC, AAMI, NIST, OECD, and IEEE related standards (Rev 7.1) [Електронний ресурс] / Orlando Lopez // LinkedIn: Log In or Sign Up. – Режим доступу: https://www.linkedin.com/pulse/ai-isoiec-related-standards-orlando-lopez-kkqqe#:~:text=,Governance

20. Ethics of Artificial Intelligence [Електронний ресурс] // https://www.unesco.org/. – Режим доступу: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=Recommendation%20on%20the%20Ethics%20of,Artificial%20Intelligence

References

1. ISO/IEC TR 24028:2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. — Geneva : International Organization for Standardization, 2020.

2. NIST AI 100-1:2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). — Gaithersburg, MD: National Institute of Standards and Technology, 2023. — 48 p. — URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

3. ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. — Geneva : International Organization for Standardization, 2023.

4. EEE Std 7000-2021. IEEE Standard Model Process for Addressing Ethical Concerns During System Design. — New York : IEEE Standards Association, 2021.

5. OECD. Recommendation of the Council on Artificial Intelligence. — Paris : Organisation for Economic Co-operation and Development, 2019 (upd. 2024). — URL: https://oecd.ai/en/ai-principles

6. European Commission. Ethics Guidelines for Trustworthy AI. — Brussels : High-Level Expert Group on Artificial Intelligence, 2019. — URL: https://digital-strategy.ec.europa.eu

7. UNESCO. Recommendation on the Ethics of Artificial Intelligence. — Paris : United Nations Educational, Scientific and Cultural Organization, 2021. — URL: https://unesdoc.unesco.org

8. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. — Geneva : International Organization for Standardization, 2023.

9. European Commission. Assessment List for Trustworthy Artificial Intelligence (ALTAI). — Brussels : High-Level Expert Group on Artificial Intelligence, 2020. — URL: https://digital-strategy.ec.europa.eu

10. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation [Elektronnyi resurs] / Natalia Díaz-Rodríguez [ta in.] // Information Fusion. – 2023. – S. 101896. – Rezhym dostupu: https://doi.org/10.1016/j.inffus.2023.101896

11. Moreno-SÃ, P. A., Del Ser, J., Van Gils, M., & Hernesniemi, J. (2025). A design framework for operationalizing trustworthy artificial intelligence in healthcare: Requirements, tradeoffs and challenges for its clinical adoption. Information Fusion, 103812.Regulation (EU) 2024/1689. Artificial Intelligence Act.

12. European Commission (2024). AI Act enters into force (povidomlennia YeK).

13. (Re)Conceptualizing trustworthy AI: A foundation for change [Elektronnyi resurs] / Christopher D. Wirz [ta in.] // Artificial Intelligence. – 2025. – T. 342. – S. 104309. – Rezhym dostupu: https://doi.org/10.1016/j.artint.2025.104309

14. Zaiarnyi, O. A., & Derkachenko, Yu. V. (2023). Deiaki osoblyvosti obrobky personalnykh danykh pry vykorystanni chat-botiv zi shtuchnym intelektom na prykladi ChatGPT. Yurydychnyi biuleten, 55-62.

15. Posykaliuk O. Shtuchnyi intelekt i personalni dani: do pytannia pro poshuk balansu interesiv [Elektronnyi resurs] / Oleh Posykaliuk // Pravo Ukrainy. – 2025. – № 2025/07. – S. 122. – Rezhym dostupu: https://doi.org/10.33498/louu-2025-07-122

16. Kobko-Odarii V. S. ROL ShTUChNOHO INTELEKTU V SUDOVII INTERPRETATsII PRAVA [Elektronnyi resurs] / V. S. Kobko-Odarii // Kyiv Law Journal. – 2023. – № 3. – S. 7–13. – Rezhym dostupu: https://doi.org/10.32782/klj/2023.3.1

17. Teremetskyi V. Artificial Intelligence as a Factor in the Digital Transformation of the Justice System [Elektronnyi resurs] / V.I. Teremetskyi, O.Ya. Kovalchuk // Forum Prava. – 2024. – T. 78, № 1. – S. 106–115. – Rezhym dostupu: https://doi.org/10.5281/zenodo.10870779

18. Habib P. E. M. AI Standards by ISO/IEC, AAMI & IEEE: Ensuring Ethical & Trustworthy AI [Elektronnyi resurs] / Prof Engr Murad Habib // LinkedIn: Log In or Sign Up. – Rezhym dostupu: http://www.linkedin.com/pulse/ai-standards-isoiec-aami-ieee-ensuring-ethical-habib-xbqif#:~:text=4,of%20false%

19. Lopez O. AI ISO/IEC, AAMI, NIST, OECD, and IEEE related standards (Rev 7.1) [Elektronnyi resurs] / Orlando Lopez // LinkedIn: Log In or Sign Up. – Rezhym dostupu: https://www.linkedin.com/pulse/ai-isoiec-related-standards-orlando-lopez-kkqqe#:~:text=,Governance

20. Ethics of Artificial Intelligence [Elektronnyi resurs] // https://www.unesco.org/. – Rezhym dostupu: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=Recommendation%20on%20the%20Ethics%20of,Artificial%20Intelligence