doi: https://doi.org/10.15407/visn2026.11.016



КРАК
Юрій Васильович —
член-кореспондент НАН України, доктор фізико-математичних наук, професор, завідувач кафедри теоретичної кібернетики Київського національного університету імені Тараса Шевченка

ЛЮДИНОЦЕНТРОВАНИЙ ШТУЧНИЙ ІНТЕЛЕКТ ТА ЙОГО ЗАСТОСУВАННЯ

Стенограма доповіді на засіданні Президії НАН України 14 жовтня 2026 року

У доповіді наголошено, що розроблення теорії та методів пояснюваного штучного інтелекту (ШІ) й людиноцентрованого підходу до створення і використання ШІ-систем є стратегічно важливим завданням для забезпечення безпеки, довіри й ефективності в багатьох сферах життєдіяльності держави. Наведено результати фундаментальних і прикладних досліджень, отримані науковцями Київського національного університету імені Тараса Шевченка та Інституту кібернетики імені В.М. Глушкова НАН України. Ці роботи спрямовано на подальший розвиток методів ШІ, що забезпечують гармонійну взаємодію людини і машини.

Шановний Анатолію Глібовичу!
Шановні члени Президії, колеги!
Дякую за можливість представити до вашої уваги результати досліджень за одним із важливих напрямів розвитку штучного інтелекту (ШІ), а саме: напрямом людиноцентрованого ШІ. У доповіді я покажу переваги й недоліки такого підходу, продемонструю конкретні можливості його застосування у важливих сферах життєдіяльності держави, де помилка у прийнятті рішень може коштувати надто дорого.

Слід зазначити, що в НАН України протягом багатьох десятиліть науковці розробляли підходи й алгоритми штучного інтелекту, закладаючи наукове підґрунтя для розвитку сучасних методів. Провідну роль у цих дослідженнях відігравав академік Віктор Михайлович Глушков. У праці «Кібернетика — любов моя», опублікованій у 1977 р., він писав: «Проблема штучного інтелекту, моделювання інтелектуальної діяльності — ось що цікавить мене в кібернетиці більш за все». Теоретичні роботи в галузі кібернетики В.М. Глушков чітко пов'язував з проблемами штучного інтелекту і говорив про медичну кібернетику, яка завдяки комплексному моделюванню допомагатиме лікарям у лікуванні хворих, особливо в критичних ситуаціях. Пам'ятаю, як у ті часи в товаристві «Знання», що розміщувалося у відомій

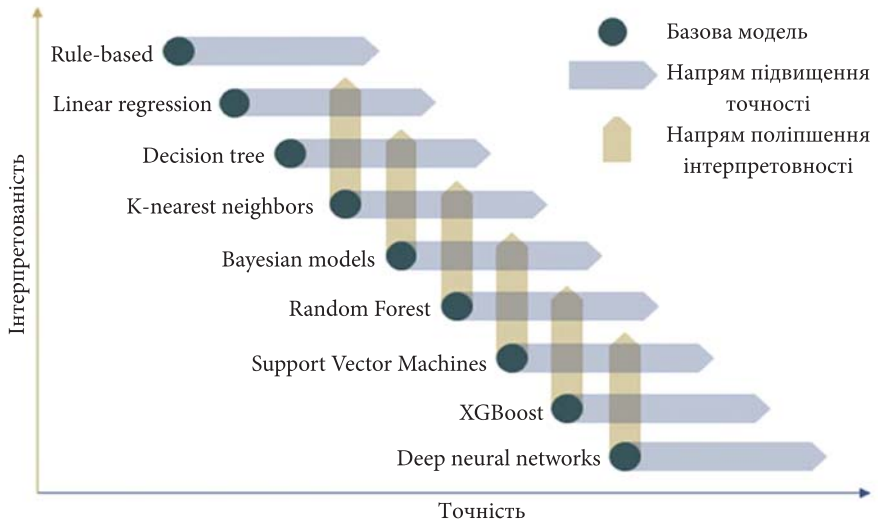


Рис. 1. Співвідношення точності та інтерпретованості моделей штучного інтелекту. З підвищенням складності й точності ШІ-моделей знижується їхня інтерпретованість, тобто людині стає дедалі складніше зрозуміти їхню логіку

будівлі на розі бульвару Тараса Шевченка та вулиці Володимирської, точилися гарячі дискусії про штучний інтелект. Нагадаю також, що в першій «Енциклопедії кібернетики» статтю про штучний розум написав академік Микола Михайлович Амосов.

Сьогодні в Україні провідним центром досліджень у галузі ШІ є Відділення інформатики НАН України, де вирішують науково-дослідницькі й організаційні питання в різних напрямах розвитку ШІ. Цю доповідь присвячено одному з таких важливих напрямів — людиноцентрованому ШІ. Об'єктом наших досліджень є процеси створення, функціонування та взаємодії складних інтелектуальних систем (нейромереж глибокого навчання, великих мовних моделей, систем комп'ютерного зору) з людиною-експертом у критичних сферах застосування, а предмет дослідження становлять математичні методи, алгоритмічні архітектури та етико-онтологічні засади, які забезпечують прозорість, пояснюваність і довіру до результатів роботи ШІ.

Що ж таке людиноцентрований штучний інтелект і чим він відрізняється від інших сучасних нейромереж глибокого навчання?

Людиноцентрований ШІ (Human-Centered Artificial Intelligence, HCAI) — це такий підхід до розроблення та впровадження систем ШІ, в якому головним пріоритетом є людина, її потреби, цінності й безпека. На відміну від традиційного підходу, часто орієнтованого на повну автоматизацію та заміну людини, HCAI зосереджується на підсиленні людських здібностей, забезпеченні контролю й довіри з боку користувача. Зокрема, згідно з цією концепцією, ШІ працює як розумний помічник, який лише допомагає людині приймати оптимальні рішення, а не замінює її повністю. Критичні рішення тут завжди залишаються за людиною.

Основні принципи HCAI — прозорість та пояснюваність, тобто користувач повинен розуміти, як саме ШІ дійшов певного висновку чи сформулював ту чи іншу рекомендацію. При цьому мають враховуватися принципи етики, справедливості та конфіденційності. Концепція людиноцентрованого ШІ ґрунтується на тріаді «людина — технологія — контекст» і принципах довіри та партнерства.

Отже, в основі наших досліджень лежить перехід від концепції ШІ-системи як «чорної скриньки», проблема якої полягає в дилемі «висока точність — низька інтерпретованість», до людиноцентрованого підходу та пояснюваного ШІ. На рис. 1 зазначену проблему традиційних ШІ-систем показано як обернену залежність точності моделей, тобто здатності моделей давати правильний результат, від їхньої інтерпретованості — здатності людини зрозуміти логіку рішення, прийнятого моделлю.

Високу точність мають складні нейромережеві архітектури, однак вони створюють істотні перешкоди для побудови моделі, зрозумілої людині. Зокрема, метод опорних векторів (Support Vector Machine) і глибокі нейронні мережі (Deep Neural Networks) оперують складними математичними перетвореннями в багатовимірних просторах, які не мають прямих аналогів у людській мові або візуальних образах. Методи Random Forest та XGBoost продукують тисячі окремих моделей, які неможливо проаналізувати вручну. Крім того, в глибокому навчанні знання розпорошені між мільйонами ваг нейронів. Тому формальна ШІ-модель стає величезною матрицею чисел, яка для ментальної моделі людини є абсолютно непрозорою. У цих методах неможливо точно сказати, який саме вхідний параметр став вирішальним для конкретного висновку.

Найвищу інтерпретованість мають прості алгоритми, які вважають когнітивно близькими до мислення людини, оскільки їхню логіку легко представити у вигляді ментальних образів або слів. Так, у методах на основі правил (Rule-Based) людина-експерт може перевірити кожне правило. Це еталон прозорості, проте такі моделі не здатні самостійно виявляти приховані складні залежності у великих масивах даних. Лінійна регресія (Linear Regression) представляє результат як зважену суму вхідних ознак, а коефіцієнти (ваги) при ознаках вказують на ступінь їхнього впливу на результат. Методи, побудовані на деревах рішень (Decision Tree), узагалі дозволяють простежити весь шлях від кореня дерева до кінцевого висновку.

Проблема «чорної скриньки» полягає в тому, що найбільш ефективні алгоритми глибокого навчання працюють як закрита система: ми знаємо, що подаємо на вхід, і отримуємо точний результат на виході, але механізм прийняття рішення залишається невідомим. Це породжує недовіру, оскільки модель може прийняти рішення на основі хибних або етично неприйнятних ознак. У критичних ситуаціях це може призвести до катастрофічних наслідків.

Тому виникає необхідність у пояснюваному штучному інтелекті (Explainable AI, XAI). Він потрібний передусім для здобуття довіри: експерт має розуміти, чому система зробила саме такий висновок. Для верифікації й аудиту важливо встановити математичний зв'язок між внутрішніми параметрами моделі та реальними параметрами. Крім того, XAI потрібен для підвищення безпеки, зменшення ризику «галюцинацій» ШІ та помилок у разі змінення контексту.

Теоретичне підґрунтя людиноцентрованого ШІ викладено в нашій монографії «Інтелектуальні інформаційні технології: людиноцентрований підхід, принципи етики та довіри» (2023). Центральним результатом у ній є онтологія довіри до інтелектуальних інформаційних технологій, побудована відповідно до ISO/IEC TR 24028. Це дозволяє перейти від суб'єктивного поняття довіри до математично і логічно вимірюваних характеристик системи.

Довіра та етика тут є критеріями якості. Довіра визначається через здатність системи мінімізувати ризики «чорної скриньки», а етика реалізується через систему обмежень, які формують прозоре середовище взаємодії. Це забезпечує керованість — право людини на остаточне рішення та можливість втручання у процес його прийняття.

Вимоги до побудови системи такі:

- прозорість, тобто відкритість архітектури для аудиту;
- керованість, зокрема наявність механізмів аварійної зупинки або ручного коригування логіки ШІ;
- конфіденційність і захист персональних даних, що особливо важливо для медичних систем аналізу;
- стратегії зменшення упередженості, щоб алгоритми не дискримінували певні групи людей через некоректні навчальні дані;
- функціональна безпека — запобігання фізичній шкоді в разі збою модуля ШІ.

Вимоги до перевірки рішень та валідації:

- пояснюваність — це ключова вимога XAI, система повинна надати логічний доказ свого висновку;
- робастність — стійкість до «шумів» або цілеспрямованих атак на дані;

• надійність і стійкість — здатність системи працювати стабільно тривалий час та відновлюватися після помилок;

• зменшення збоїв апаратного забезпечення — перевірка того, як програмний ШІ взаємодіє із «залізом»;

• тестування та оцінка — обов'язкова наявність метрик, які підтверджують якість рішення;

• застосовність — перевірка того, чи відповідає рішення конкретному контексту.

Отже, використання стандарту ISO/IEC TR 24028 дозволяє створювати не просто «розумні» алгоритми, а соціально відповідальні технології, які мають юридичне та логічне обґрунтування, що є критичним для їх впровадження в державний та оборонний сектори.

Розглянемо архітектуру інтелектуального інтерфейсу, який є містком між штучним інтелектом і людиною-експертом. Ключовим результатом є математично обґрунтована схема перетворення, що дозволяє інтерпретувати внутрішні стани глибокої нейронної мережі. У машинному вимірі процес починається з наборів даних, які проходять через нейромережу глибокого навчання. На етапі вибору вбудовування (embedding) формується формальна модель. Центральною частиною (ядром) XAI є інтерфейс наближення, де відбувається головна математична операція синхронізації. У людському вимірі результат проєктується на ментальну модель, зрозумілу фахівцю і засновану на візуальних образах та логіці.

Під час синхронізації виникає кілька проблем. Передусім це когнітивний розрив: людина не може безпосередньо сприймати тензорні простори та мільйони ваг нейромережі. Інша проблема — приховані залежності: мережа знаходить нелінійні патерни, які часто не мають назви в людській мові. Нарешті, математична складність формальної моделі робить її закритою («чорна скринька»), що унеможливлює швидку перевірку логіки рішення.

Існує глобальний дефіцит математичних методів, які поєднують високу точність нейромережевих підходів із прямою математичною синхронізацією ментальних образів. Серед сучасних підходів до пояснюваності виділяють апріорні (ante-hoc) та апостеріорні (post-hoc) методи.

Апріорні методи роблять модель зрозумілою вже на етапі її проєктування. Такі моделі називають «скляними скриньками» (glass-box), або внутрішньо інтерпретованими. Пояснення не є додатком до моделі: логіка прийняття рішення закладена в її архітектуру. Наприклад, дерева рішень описують ієрархію правил «якщо — то», баєсівські списки правил містять набори логічних умов, які легко прочитати людині, а узагальнені адитивні моделі дозволяють моделювати нелінійні зв'язки, зберігаючи можливість візуалізувати внесок кожної ознаки окремо.

Апостеріорні методи застосовують до вже навчених моделей, зазвичай складних «чорних скриньок» — нейронних мереж або випадкових лісів, щоб зрозуміти, як вони працюють. Модель залишається незмінною, а окремий алгоритм аналізує її входи та виходи і створює наближене пояснення. Зокрема, він визначає внесок кожної ознаки у фінальний прогноз, знаходить мінімальний набір умов, за яких прогноз не зміниться, або показує найменшу зміну вхідних даних, яка змінила б рішення. До цих підходів належать LIME, SHAP і Grad-CAM.

Стратегія нашого наукового напряму є комплексним методологічним ланцюжком, який дозволяє трансформувати непрозорі алгоритми глибокого навчання в зрозумілі людині інтелектуальні системи. Вона складається з трьох ключових етапів, а саме:

1) математичне наближення — оскільки ШІ і людина оперують різними ознаками, на етапі математичного наближення необхідно відобразити абстрактні математичні вектори нейромережі на конкретні образи, які мають сенс для фахівця. Це дозволяє встановити, яким реальним фізичним або логічним параметрам відповідають внутрішні ознаки моделі;

2) структурна декомпозиція полягає у відмові від монолітної архітектури на користь ансамблевого та ієрархічного підходів. Модель розбивається на логічні блоки, кожен із яких відповідає за певну зрозумілу категорію знань, наприклад геометрію об'єкта, його текстуру або динаміку змін. Це робить систему прозо-

рою не лише на виході, а й на етапі проміжних обчислень;

3) архітектурна дистиляція — це процес стиснення знань великої й складної моделі-вчителя в компактну, логічно впорядковану модель-учня. Мета полягає в тому, щоб зберегти високу точність глибокого навчання, але реалізувати її в архітектурі, де кожен шар і кожен нейрон мають чітке логічне тлумачення та забезпечують повну простежуваність.

Кінцевим результатом цієї триступеневої стратегії має стати вбудована прозорість ШІ.

Наведемо математичний апарат зв'язку між формальними й ментальними моделями. Запропонований підхід ґрунтується на мінімізації нев'язки — знаходженні оптимальної матриці переходу T між простором ознак нейромережі A та простором ознак експерта B. Формальна модель A описує m об'єктів k ознаками нейромережі, а ментальна модель B — ті самі об'єкти l ознаками, зрозумілими людині. Матриця T мінімізує розбіжність між AT і B (норма Фробеніуса). Розв'язок дає псевдообернена матриця Мура—Пенроуза. Сингулярний розклад дозволяє виконати обчислення і для виродженої матриці A. Регуляризація Тихонова зменшує вплив малих сингулярних чисел, щоб пояснення моделі не змінювалося радикально за незначних змін вхідних даних. Новий об'єкт пояснюємо множенням його ознак на матрицю T.

Як працює такий підхід, покажемо на прикладі ЕКГ. Він демонструє, як абстрактні числа нейромережі перетворюються на зрозумілу лікарю мову. На етапі навчання взято п'ять записів ЕКГ. Для кожного запису нейромережа виділила три числові ознаки, які самі по собі нічого не кажуть лікарю. Для тих самих п'яти записів лікар-експерт визначив два зрозумілі параметри: тривалість комплексу QRS у мілісекундах та частоту серцевих скорочень в ударах за хвилину.

За формулою з псевдообернено матрицею обчислюємо матрицю переходу T, яка найкраще відображає числові ознаки нейромережі на показники, обрані лікарем. У наведеному прикладі відносна похибка наближення становить близько 0,005, тобто 0,5 %. Це демонструє математичний зв'язок між внутрішніми ознаками ШІ та знаннями лікаря.

Тепер, маючи такий «перекладач», можемо перейти до пояснення нового запису ЕКГ. Нейромережа видає три формальні ознаки: 1,4; 0,6; 1,0. Множення цього вектора на матрицю T дає тривалість QRS приблизно 86,9 мс і частоту серцевих скорочень близько 66,1 уд./хв. Інтерфейс візуальної аналітики показує зв'язки між формальними ознаками та цими показниками. Лікар отримує кількісні характеристики й бачить, які компоненти моделі за них відповідають. Саме в цьому полягає перетворення «чорної скриньки» на «білу».

Докладніше зупинюся на стратегії структурної декомпозиції. Замість однієї складної нейромережі, яка намагається вивчити все одразу, будуємо ансамбль логічно розділених вузлів. Кожен вузол спеціалізується на вилученні ознак, що мають чітке фізичне або медичне тлумачення для людини. Послідовність роботи наведено на рис. 2: вхідні дані (МРТ, ЕКГ або зображення з БпЛА) паралельно обробляються спеціалізованими вузлами, після чого експерт перевіряє вилучені ознаки, а результати об'єднуються в консолідоване рішення.

Вузол А відповідає за геометрію та контури, тобто за виявлення просторової структури об'єкта. У медицині це форма органів, у завданнях для БпЛА — обриси будівель чи доріг. Математичну основу становлять оператори градієнта, згорткові фільтри для виявлення країв і дескриптори форми. Вхідний сигнал відображається у простір геометричних примітивів (лінії, кути, криві), де результат роботи нейронів прямо корелює з геометричним параметром об'єкта.

Вузол Б аналізує текстуру та щільність — внутрішні властивості матеріалу або тканини. Для МРТ це важливо для розрізнення здорової й патологічної тканини, наприклад міокарда, за її характеристиками; для БпЛА — для ідентифікації типу поверхні: ліс, вода, бетон. Тут використовують статистичні ознаки другого порядку, зокрема матриці суміжності пікселів, а також спектральний аналіз — перетворення Фур'є або вейвлети — для оцінювання частотних характеристик зображення. Результати

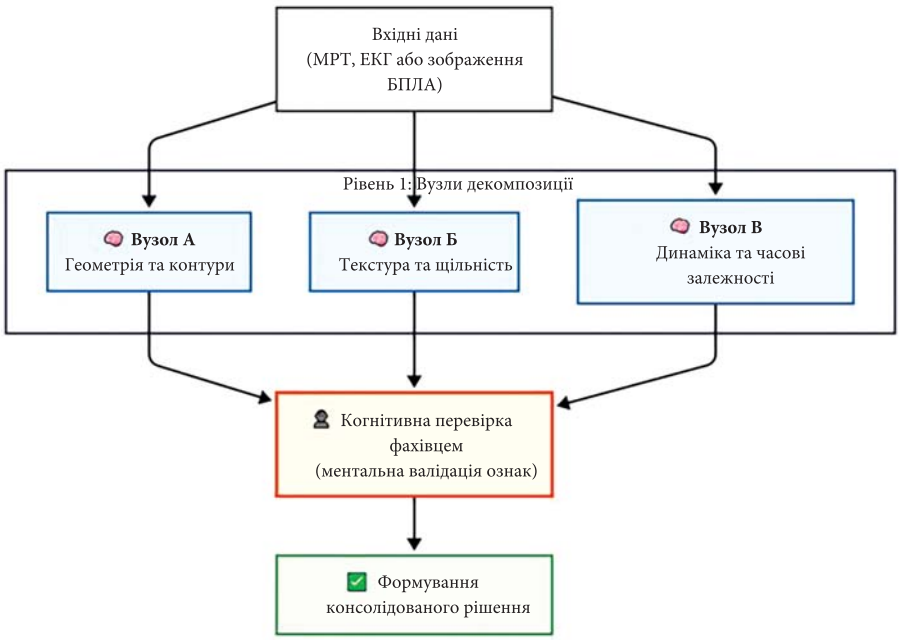


Рис. 2. Структурна декомпозиція та ансамблевий підхід до формування зрозумілих рішень

формують своєрідний «підпис» матеріалу в ментальній моделі людини.

Вузол В відповідає за динаміку й часові залежності. Це ключовий вузол, де для ЕКГ аналізують динаміку серцебиття, а для БпЛА — відеопотік з дрона. Він дає можливість аналізувати процеси, а не лише статичні зображення. Для цього використовують рекурентні структури LSTM і GRU або оптичний потік. Математично це моделювання часових рядів, у якому внутрішні стани моделі відповідають фазам фізіологічного чи механічного процесу, наприклад фазам скорочення шлуночків серця.

На завершальному етапі ансамблевий підхід формує єдину інтерпретовану архітектуру. Каркасом є людська логіка: послідовність кроків прийняття рішення задається людиною. Математично це означає накладання обмежень на структуру зв'язків між шарами нейромережі. Кожен вузол ансамблю має бути корельованим із ментальною ознакою, щоб можна було обчислити внесок саме геометрії або динаміки у фінальний висновок. Далі знання ансамблю дистилюються в одну нейромережу з прозорою структурою шарів. Кожен шар виконує роль одного з вузлів декомпозиції, що забезпечує простежуваність прийняття рішення від входу до виходу.

Отже, структурна декомпозиція дозволяє замінити «магію» нейромереж строгим інженерним підходом, за якого ШІ працює відповідно до людської логіки, зберігаючи обчислювальну потужність глибокого навчання.

Дистиляція в інтерпретовані архітектури є завершальним етапом перетворення складних моделей на «білі скриньки». Схему цього процесу наведено на рис. 3. Його мета — поєднати точність на рівні найкращих сучасних алгоритмів із повною простежуваністю логіки прийняття рішень.

Перший блок — логічно декомпозований ансамбль. Задача розбивається на незалежні підзадачі, а кожна модель ансамблю спеціалізується на конкретній, зрозумілій людині ознаці. Це дає змогу пов'язати частини інтелектуальної системи з реальними фізичними або семантичними категоріями. Математично цільову функцію представляють як композицію функцій, кожна з яких є окремою нейромережевою моделлю, що відображає вхідні дані у простір специфічних ознак, корельованих із ментальними образами експерта.

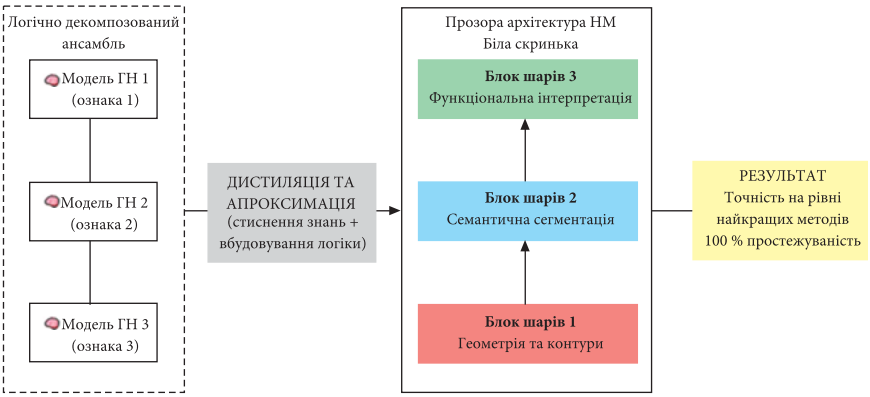


Рис. 3. Дистиляція логічно декомпозованого ансамблю в інтерпретовану нейромережеву архітектуру;
ГН — глибоке навчання;
НМ — нейромережа

Другий блок — дистиляція та апроксимація. Інформація й закономірності, накопичені складним ансамблем-вчителем, стискаються і переносяться в одну компактну архітектуру-учня. Головна інновація полягає у вбудовуванні логіки: модель-учень структурує шари відповідно до логічних етапів декомпозиції. Методи дистиляції знань використовують, щоб мінімізувати комбіновану функцію втрат між розподілами ймовірностей ансамблю та нової прозорої моделі. Апроксимація слугує засобом знаходження матриць переходу між неструктурованими вагами вчителя та структурованими блоками учня.

Третій блок — прозора архітектура нейромережі, що складається зі спеціалізованих блоків шарів. Перший з них відповідає за геометрію й контури: вилучає низькорівневі ознаки — краї, кути, форми. Його математичною основою є згорткові фільтри із заданими або регуляризованими вагами, які реалізують оператори диференціювання зображення.

Другий блок шарів забезпечує семантичну сегментацію. На основі геометрії ШІ він виділяє змістовні зони — наприклад, лівий шлуночок у кардіології або конкретні будівлі на знімку з БпЛА. Математично це формування масок сегментації, обчислення просторових дескрипторів і відображення геометричних ознак у класи об'єктів.

Третій блок шарів здійснює функціональну інтерпретацію, тобто формує логічний висновок. Використовуючи вилучені ознаки та вбудовану базу знань, він надає пояснення: чому це аритмія або чому це об'єкт загрози. Тут застосовують лінійні шари, що працюють як «м'яка логіка», а ваги нейронів відображають важливість ознаки для фінального висновку.

Результатом має стати баланс між високою точністю й повною простежуваністю — можливістю аудиту кожного кроку: від того, яку лінію виявив перший шар, до її впливу на кінцеве рішення. У представленій концепції йдеться про узгодженість формального висновку з логічним поясненням, яка забезпечується цілісністю ланцюжка дистиляції. Така стратегія спрямована на перетворення ШІ на доказовий інструмент.

За браком часу перелічу лише деякі важливі прикладні застосування людиноцентрованого штучного інтелекту. Перший приклад — кардіологія та аналіз МРТ, де поєднано багатоетапну сегментацію, каскадну класифікацію й візуальну інтерпретацію. Причому візуальна інтерпретація тут потрібна для пояснення висновку системи. Структуровані результати демонструють анатомічні й функціональні показники, зокрема зміни товщини стінок міокарда, параметри їхнього руху та геометрію порожнин. Лікар бачить наочні карти подібності, контурні маски й кількісні маркери, які може враховувати під час ухвалення рішення.

Інший приклад — виявлення аритмій за ЕКГ із забезпеченням прозорості аналізу сигналу. ШІ розшифровує ЕКГ та пропонує варіант діагнозу, вказуючи конкретну причину висновку. При цьому є нагальна потреба створити наці-

ональний набір даних ЕКГ із залученням профільних інститутів НАМН України та з урахуванням етичних принципів і норм законодавства.

Для великих мовних моделей представлено підхід до математичного обґрунтування та логічного виділення ознак в аналізі соціальних мереж. Зокрема, розроблено методи виявлення політичної пропаганди й кібербулінгу, що важливо для інформаційної безпеки держави та особистої безпеки користувачів Інтернету.

Наступна група прикладів стосується пояснюваного комп'ютерного зору для автономних систем БпЛА. Загальна ідея — надійно розпізнавати візуально подібні об'єкти на знімках, замінивши одну «плоску» модель ієрархічним каскадом. Кожен його рівень уточнює клас об'єкта від загального до конкретного. Структура каскаду будується автоматично, без залучення експертів, за гібридними відстанями між класами, які поєднують геометрію ознак і реальні помилки класифікатора, та з використанням агломеративної кластеризації. Підтверджено, що підхід зберігає високу точність, зокрема й під час перенесення на інший набір даних.

Систему адаптовано до реальних умов роботи на борту БпЛА: поріг розсічення дендрограми враховує допустиму затримку конкретної обчислювальної платформи. Для отримання спеціального ембедингу з проміжних шарів використовують модель глибокого навчання YOLO. Ці ознаки пов'язують зі зрозумілими характеристиками зображення, а далі аналізують за допомогою компактної табличної моделі трансформера.

Для погодостійкої візуальної локалізації БпЛА створюють синтетичні зображення з туманом, снігом і нічним освітленням, перевіряючи збереження геометрії будівель. За внутрішніми шарами моделі YOLO отримують ембедінги й формують компактний опис сцени для зіставлення кадру з БпЛА із супутниковою картою. Для міської забудови в умовах синтетичного поєднання туману, снігу та нічного освітлення запропонований підхід зберіг 65,8 % свого рівня якості локалізації в чистих умовах, тоді як сама модель YOLO11 — лише 46,6 %.

Ще одне застосування — візуальна корекція інерціальної навігації за відсутності GPS. Підхід допомагає коригувати накопичену похибку завдяки візуальній ідентифікації орієнтирів. Перед польотом на супутникових знімках сегментують об'єкти й формують їхні ембедінги. Ембедінги та координати найвиразніших об'єктів зберігають у банку орієнтирів, на основі якого планують маршрут. Під час польоту система перевіряє, чи є серед об'єктів у кадрі наступний запланований орієнтир, і передає його координати для корекції похибки лише після підтвердження збігу.

І нарешті — пояснюване навчання з підкріпленням і планування майбутніх систем розумного транспорту. Ефективне та стале планування ускладнюється застарілими моделями управління рухом, які не враховують реальну динаміку. Це призводить до заторів і значного впливу на довкілля. Для інтелектуального регулювання транспортних потоків запропоновано адаптивну каскадну архітектуру кластеризації. Процес починається зі збирання даних міської транспортної мережі, з яких вилучають ознаки в межах окремих часових вікон. Далі механізм зваженого голосування обирає оптимальну стратегію кластеризації на основі внутрішніх характеристик даних. Це забезпечує ідентифікацію різних типів транспортних потоків. Розроблену модель перевірено за допомогою високоточного моделювання транспортної мережі міста і впроваджено в м. Хмельницький.

Отже, головний результат синтезу концепцій полягає в переході від апостеріорних пояснень, коли аналізують уже навчену «чорну скриньку» зовнішніми методами, до вбудованої прозорості — створення архітектур, зрозумілих за своєю побудовою.

Ядром цього синтезу є стратегія «білої скриньки», яка спирається на чотири складові. Етична онтологія передбачає врахування положень ISO/IEC TR 24028 та морально-правових норм безпосередньо в логіці алгоритму. Математична апроксимація використовує апарат матриць переходу, щоб пов'язати математичні

оператори із семантичними аналогами в ментальній моделі людини. Структурна декомпозиція означає проєктування мережі як ансамблю прозорих вузлів, що відповідають за геометрію, текстуру та динаміку. Мотиваційна прозорість забезпечує розуміння цілей і стимулів моделі: людина має бачити, на що налаштована функція винагороди ШІ, особливо в системах навчання з підкріпленням.

Виокремимо три стратегічні вектори подальших досліджень. Перший — самодокументовані нейромережі. Це системи, які під час кожного висновку, наприклад щодо діагнозу або маневру БпЛА, автоматично генеруватимуть доказову базу — текстовий чи візуальний звіт про чинники, що привели до рішення, та їхню ієрархію. У перспективі це допоможе автоматизувати юридичний і медичний аудит.

Другий вектор — дистиляція знань у «ментально коректні» топології. Подальші дослідження спрямовуватимемо на те, щоб структура зв'язків нейромережі відтворювала ієрархічну логіку людського мислення. Перспективним є створення «когнітивних двійників» експертів — перенесення досвіду лікарів чи аналітиків у структуру ШІ як системи логічних взаємозв'язків.

Третій вектор — гармонійна взаємодія людини та машини. Система має працювати в такому темпі й з такими образами, які узгоджуються з когнітивними можливостями людини. Це допоможе подолати розрив верифікації: в умовах зростання складності систем ШІ XAI-інтерфейси дозволятимуть людині залишатися в контурі управління, зберігаючи обчислювальну потужність машини.

Пояснюваність ми розглядаємо як засіб здобуття безпеки й довіри зі збереженням точності. Подальші дослідження плануємо зосередити на створенні інтелектуальних систем, у яких математична строгість глибокого навчання поєднується з прозорістю людської мови та візуальних образів. Це перетворює ШІ на контрольованого партнера людини.

Завершуючи доповідь, зверну увагу на три групи отриманих результатів. У теоретико-етичному аспекті сформовано онтологію довіри та систему «етичних фільтрів». У математично-архітектурному — розроблено методи апроксимації з використанням матриць переходу та декомпозиції формальних моделей у ментальні. У практично-валідаційному — підтверджено ефективність підходів у критичних сферах, зокрема медицині, використанні БпЛА й екології.

Людиноцентрований штучний інтелект — це безпека через розуміння. Це технологія, яка робить людину сильнішою, не позбавляючи її контролю.

Дякую за увагу!

За матеріалами засідання підготувала О.О. Мележик

Iurii V. Krak
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: https://orcid.org/0000-0002-8043-0785

HUMAN-CENTERED ARTIFICIAL INTELLIGENCE AND ITS APPLICATIONS

Transcript of scientific report at the meeting of the Presidium of NAS of Ukraine, October 14, 2026

The report emphasizes that the development of the theory and methods of explainable artificial intelligence (AI) and a human-centered approach to the creation and use of AI systems is a strategically important task for ensuring security, trust, and efficiency in many areas of the state's life. The results of fundamental and applied research obtained by scientists of the Taras Shevchenko National University of Kyiv and the V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine are presented. These works are aimed at further developing AI methods that ensure harmonious interaction between man and machine.

Cite this article: Krak I.V. Human-centered artificial intelligence and its applications (transcript of scientific report at the meeting of the Presidium of NAS of Ukraine, October 14, 2026). *Visn. Nac. Akad. Nauk Ukr*. 2026. (11): 16—24. https://doi.org/10.15407/visn2026.11.016