КІБЕРНЕТИКА

УДК 004.89

В.С. ХАРЧЕНКО
Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна, e-mail*: v.kharchenko@csn.khai.edu.*

С.В. ЯКОВЛЕВ
Харківський національний університет ім. В.Н. Каразіна, Харків, Україна, e-mail: *s.yakovlev@karazin.ua.*

О.Ю. ВЕПРИЦЬКА
Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна, e-mail: *o.veprytska@csn.khai.edu.*

Г.В. ФЕСЕНКО
Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна, e-mail: *h.fesenko@csn.khai.edu.*

ПОЯСНЕ́ННИЙ ШТУЧНИЙ ІНТЕЛЕКТ ЯК СЕРВІС: АЛГОРИТМИ ОЦІНЮВАННЯ ХАРАКТЕРИСТИК

Анотація. Досліджено методи оцінювання поясненного штучного інтелекту як сервісу (XAIaaS). Проаналізовано підходи до оцінювання систем поясненного штучного інтелекту. Метою дослідження є розроблення підходу та послідовності оцінювання якості систем поясненного штучного інтелекту в контексті моделі XAIaaS, що ґрунтується на характеристико-орієнтованій методології. Запропоновано багаторівневу схему оцінювання, яка передбачає поетапний аналіз, тестування та корекцію на підставі визначених ключових характеристик першого рівня ієрархії моделі якості, що дає змогу здійснювати цільове, структуроване та порівняльне оцінювання окремих властивостей системи в цілому. Продемонстровано застосування запропонованої методики оцінювання на практичному прикладі.

Ключові слова: штучний інтелект, XAI як послуга, оцінювання поясненного штучного інтелекту, профілювання вимог.

ВСТУП

Стрімкий розвиток та інтеграція систем штучного інтелекту (artificial intelligence system, AIS) у критично важливі галузі, як-от: охорона здоров'я, фінанси, оборона та безпека, енерегетика і державне управління, супроводжуються посиленням вимог до прозорості та надійності таких систем. Серед AIS значний внесок зробили моделі глибокого навчання (DL) [1], які демонструють вражаючі результати в задачах класифікації, прогнозування, виявлення аномалій, генерації рекомендацій тощо. Проте підвищення точності цих моделей відбувається переважно за рахунок збільшення їхньої складності, що, зі свого боку, ускладнює пояснення логіки їхнього функціонування. У низці випадків це призводить до рішень, які суперечать очікуваній логічній поведінці в конкретному контексті, водночас система не надає жодного обґрунтування причин такого вибору. Ця непрозорість значно ускладнює довіру до AI, унеможливлює цільову оптимізацію моделей та обмежує їхнє використання у реальних умовах.

Без надійних пояснень, що адекватно відображають внутрішні процеси AI, користувачі сприймають такі системи як непередбачувані та ненадійні.