ІСТОРИЧНИЙ ФАКУЛЬТЕТ

Скочипець Денис
Науковий керівник – доц. Грушко Віктор

ЕСКАЛАЦІЯ РИЗИКІВ В УМОВАХ ШВИДКОГО РОЗВИТКУ ШТУЧНОГО ІНТЕЛЕКТУ

У статті здійснено теоретичний аналіз ризиків, які обумовлені впливом прискореного розвитку науково-технічного прогресу, зокрема штучного інтелекту в умовах сучасного складноорганізованого суспільства. На основі концептуального доробку Ніка Бострома – ідей інтелектуального вибуху, інструментальної конвергенції, проблеми контролю та положень ризикології систематизовано основні класи загроз, серед яких виділені: соціально-економічні, політичні, інформаційні та екзистенційні. Досліджено ризики, що безпосередньо зачіпають повсякденне життя пересічних людей. Обґрунтовано необхідність комплексного управління ризиками, які здатний породжувати штучний інтелект. Здійснений аналіз ризиків які виявляють тенденцію до зростання при відсутності ефективного контролю над ШІ. Визначено як одну з найгостріших загроз для людства проблемність забезпечення контролю над діяльністю ШІ, що може спричинювати дестабілізацію функціонування суспільних інститутів і нести небезпеку для кожної окремо взятої людини.

Ключові слова: штучний інтелект, технологічний ризик, екзистенційний ризик, системний ризик, інструментальна конвергенція, проблема контролю, управління ризиками.

Постановка проблеми. Особливістю сучасного етапу зростання ризиків під впливом останньої хвилі науково-технічних перетворень (у порівнянні з тими які змінювали життя людей на етапах індустріальних революцій) є те, що раніше техногенні загрози мали певну локалізацію. Штучний інтелект руйнує цю логіку. Він не має конкретної адреси, не має єдиного вузла вразливості – і саме тому традиційні інструменти управління ризиками починають давати збій.

Сучасне техногенне суспільство на нинішньому етапі свого розвитку опинилося в ситуації, коли найбільш проникна з усіх технологій глибоко вбудована у найрізноманітніші сфери життя соціуму. За прогнозами PricewaterhouseCoopers внесок ШІ у глобальний ВВП до 2030 року може сягнути 15,7 трлн дол. США [1, с. 57]. ШІ вже сьогодні трансформує умови людського існування способами, які не є ані повністю зрозумілими, ані достатньо підконтрольними людям.

У статті проаналізовано ключові класи ризиків, що породжуються технологіями штучного інтелекту на основі концептуальних підходів шведського філософ-трансгуманіста при Оксфрдському університеті, засновника Інституту майбутнього людства Ніка Бострома.

Метою даної статті є аналіз встановлення потенційних ризиків, зумовлених використанням ШІ, та визначення підходів до управління ними в умовах сучасного складноорганізованого суспільства.

Аналіз досліджень та публікацій. Багаторівневий характер ризиків ШІ обумовлює необхідність досліджень комплексу технічних, правових, філософських та інших аспектів цієї проблематики. Зокрема, дослідження OpenAI (Хейді Хлааф, Памела Мішкін) заклали методологічну базу для ідентифікації та класифікації небезпек ШІ, тоді як фундаментальна праця Ніка Бострома розглядає стратегічні загрози виникнення суперінтелекту, що виходять далеко за межі питань кібербезпеки. В українському науковому доробку дана проблема вивчалась науковцями: Генадієм Андрощуком, який акцентує на кримінальних загрозах і викликах для інтелектуальної власності; Володимиром Скіцько, Павлом Складанним, які інтегрують ризики ШІ у загальну систему кібербезпеки на основі стандарту ISO 31000:2018: Андрієм Колесніковим та Ольгою Карапетян, чий доробок присвячений визначенню пріоритів у вітчизняному законодавчому регулюванні, зокрема у сфері етики та захисту персональних даних.

Виклад основного матеріалу. Нік Бостром у своїй роботі «Суперінтелект» здійснив філософське осмислення та аналіз питання можливих наслідків для людства створення ним чогось розумнішого, без гарантій того, що воно буде поділяти ті ж цінності, що й люди. Він вважає, що якщо виникне система ШІ, здатна покращувати власну архітектуру, то кожне покращення збільшуватиме здатність цього нелюдського розуму до наступного вдосконалення себе – і цей процес може пришвидшитись до такої міри, що людський контроль за ним стане неможливим. З часом спроможності штучного розуму одержать можливість досягнути таких розумових здібностей, які значно перевищуватимуть весь інтелектуальний потенціал людства. Після цього реальність зміниться до невпізнаваності – і це при тому, що розумові здібності систем ШІ швидше за все продовуватимуть зростати [2, с. 83]. Н. Бостром стверджує, що ймовірність цього достатньо висока, а масштаб наслідків буде велитенський.

Н. Бостром розвиває концепцію інструментальної конвергенції, в якій обгрунтовує, що більшість стратегічних цілей породжують проміжні, інструментальні прагнення: самозбереження, накопичення ресурсів, запобігання перепрограмуванню [2, с. 135]. Це означає, що небезпека від надпотужного ШІ є структурною і при цьому не залежить від злого умислу його творців. Система, яка не хоче нічого «поганого» у людському розумінні, все одно може становити екзистенційну загрозу просто тому, що її інструментальні прагнення можуть вступити в конфлікт із потребами людей у ресурсах, у свободі чи просто у виживанні. Забезпечення контролю, як стверджує Н. Бостром, і є засобом недопущення потенційних

загроз. Суто технічна проблема безпечного для людей програмування системи набуває на сьогодні першочергової актуальності. Зокрема, В. Скіцько зауважує, що вже сьогоднішні системи ШІ сприймаються суспільством як своєрідні «магічні чорні скриньки», чиї рішення неможливо ані пояснити, ані ефективно запобігати шкідливим для людей діям з їх боку [4, с. 7]. Ці обставини роблять проблему контролю актуальною вже зараз, а не лише в гіпотетичному майбутньому, коли суперінтелект досягне критично великої потуги.

Надгострою проблемою може також стати зростання технологічного безробіття. У той час, коли промислові революції XVIII-XIX століть витісняли переважно фізичну, некваліфіковану працю, нинішня техніко-технологічна трансформація вражатиме значно ширше коло працівників – включно з тими, хто спеціалізується на розумовій роботі [6, с. 13].

Ще одним з ризиків від експансії ШІ є поглиблення нерівності. За розрахунками аналітиків, Китай очікує зростання ВВП від ШІ на 26%, Північна Америка – на 14%, тоді як більшість країн Глобального Півдня – в рази менше [1, с. 57]. Усередині окремих суспільств картина схожа: ті, хто вже має доступ до технологій, освіти й капіталу, отримають від ШІ непропорційно більше. Ті ж, хто його не має, – непропорційно більше ризикують. Дослідники OpenAI фіксують, що нерівний доступ до інструментів ШІ «збільшує нерівність і скорочує економічне зростання» [6, с. 13]. Ще один ризик – алгоритмічна дискримінація. ШІ вирішуючи, чи надавати тій чи іншій людині кредит, чи запрошувати когось на співбесіду, чи визначати комусь страховий тариф, опиратиметься на упередження, що були в нього закладені. Хейді Хлааф, керівник науковців з питань ШІ в Інституті AI Now, класифікує дискримінаційну шкоду як критичну [6, с. 11–12]. При цьому жертви подібних дискримінаційних рішень можуть і не знати, що воно прийнято алгоритмом і за якою логікою.

Дипфейки (підробні зображення, відео) останнім часом стали найбільш публічно обговорюваною технологічною загрозою. Зокрема, науковий співробітник НДІ інтелектуальної власності НАПрН Г. Андрощук, посилаючись на дослідження UCL, звертає увагу, що там, де аудіо- та відеоімперсоналізація посідає перше місце, поширення дипфейків здатне спричинити «широку недовіру до відеоматеріалів як таких» [1, с. 64]. Таким чином, відеофальшивки спричинюють руйнування можливості достовірної верифікації – стан, коли будь-яке відео може бути оголошене фейком, а будь-який фейк може претендувати на автентичність. Правосуддя, виборчі процеси, довіра до ЗМІ, тощо ґрунтуються на припущенні, що докази можна перевірити. ШІ підриває довіру ледь не до кожного доказу.

Паралельно з цим алгоритми ШІ у соціальних мережах здатні визначати пріоритетність контенту, наповнюючи його однобоко поданим змістом задля маніпулювання емоціями і поведінкою користувачів [4, с. 12]. У цьому випадку сама логіка, закладена у ці алгоритми, систематично просуває контент, що викликає тривогу, обурення і поляризацію, – просто тому, що саме такий контент найефективніше утримує увагу. Результат – «інформаційні бульбашки» і поглиблення суспільних розколів. Окремо стоїть питання автономних збройних структур. ШІ

може бути застосований для створення роботизованих платформ, що самостійно зможуть визначати та атакувати цілі без участі людей-операторів [4, с. 12]. Це актуалізує питання як юридичної, так і моральної відповідальності за рішення, яке прийматиме алгоритм: хто має відповідати – виробник, держава, що його розгорнула, сам алгоритм? Сучасне право не має відповіді на жодне з цих питань.

Одна з найбільших недооцінених загроз від ШІ пов'язана з тим, що системи ШІ просто брешуть – впевнено, переконливо і систематично. Це не технічний збій з боку ШІ, який можна виправити. Це структурна властивість систем, що навчаються генерувати правдоподібний текст, а не верифікувати факти [4, с. 10]. Водночас, значно більшою є загроза цілеспрямованої дезінформації. Г. Андрощук фіксує, що ШІ здатний автоматично продукувати безліч версій певного контенту з різних удаваних джерел, щоб підвищити його позірну достовірність [1, с. 67]. Фактично, це виробництво брехні у промислових масштабах. Середньостатистична людина не має ані часу, ані інструментів, щоб перевіряти кожне твердження, з яким вона стикається в інформаційному просторі. Х. Хлааф вводить поняття «помилки з наслідками» – ситуації, коли системи ШІ спонукають людей чи інститути ухвалювати хибні рішення, що матимуть несприятливий вплив на їх подальше життя [6, с. 11]. Це може бути невірний медичний діагноз, хибна юридична порада, помилкова фінансова рекомендація –будь-яка ситуація, де людина довірилась системі, що видавала себе за компетентну. І що важливо: відповідальність у такому разі виявляється розмитою між розробником, постачальником і кінцевим користувачем.

Н. Бостром відзначає: «На якомусь етапі зростання інтелектуальних можливостей системи перетнуть межу, яку людство визначило, як точку «неповернення». Це такий рівень спроможностей, понад який покращення відбуватимуться без втручання людей – завдяки діяльності самих систем» [2, с. 83]. Ключова проблема сконцентрована тут у незворотності того, що може відбутись. Загроза не обов'язково полягає у тому, що ШІ «захоче» припинити існування людства, чи повністю побавити його суб'єктності у той чи інший спосіб. Загроза полягає у тому, що цілі надпотужної системи можуть бути просто байдужими до людського існування – так само, як люди, прокладаючи дорогу, не «бажають» знищувати мурашник, але знищують його, бо він заважає. «Загалом перший суперінтелект може визначати майбутнє земного життя, мати неантропоморфні цілі, і, найімовірніше, володітиме інструментальними можливостями, щоб розгорнути захоплення ресурсів» [2, с. 142]. Коли два суб'єкти з різними цілями та різними можливостями впливу займають один простір, то слабша сторона може залишитись ні з чим.

Дослідник у сфері кібербезпеки В. Скіцько фіксує, що вже сьогоднішні системи ШІ здатні здійснювати атаки на інформаційне забезпечення військових навчань [4, с. 9]. І якщо такі маніпуляції можливі з нинішніми, відносно примітивними системами, то, очевидно, що з потужнішими проблеми виявляться помітно більшими. Прибічники призупинення подальшого удосконалення систем ШІ стверджують: «Потужні системи штучного інтелекту слід

розробляти лише у тому випадку, якщо люди впевнені, що ефект від них буде позитивним, а ризики керованими» [4, с. 6].

Водночас, на нормативно-правовому рівні вже здійснюються певні кроки для створення запобіжників для недопущення розвитку небезпечних сценаріїв [1, с. 58–59]. Але між появою законів і реальним управлінням ризиками пролягає довгий шлях до втілення. Проблема не лише в тому, що законодавство запізнюється за розвитком технологій – більшість правових систем побудовані навколо відповідальності осіб. Коли ж шкоду заподіює алгоритм, ця логіка руйнується. Водночас, Х. Хлааф звертає увагу на «неоднозначну правову відповідальність для творців моделей, клієнтів і кінцевих користувачів» [6, с. 13].

В Україні на сьогодні відсутнє спеціалізоване законодавство у сфері ШІ. Досі будь-які норми про захист прав людини у взаємодії з алгоритмами залишаються здебільшого декларативними. Очевидно, що відповідальність розробників не повинна обмежуватись дотриманням мінімальних стандартів. На організаційному рівні В. Скіцько наголошує на необхідності об’єднання зусиль науковців, розробників, держави та міжнародної спільноти – адже жоден з цих суб’єктів поодинці не має ані ресурсів, ані повноважень для адекватної відповіді [4, с. 15]. Г. Андрощук, своєю чергою, звертає увагу на роль освітніх кампаній у формуванні обізнаності суспільства про ризики ШІ [1, с. 68].

Висновки. Ризики ШІ є новими, багатоаспектними та здатними до самопосилення, що виключає можливість їхнього подолання у межах традиційних логік техногенних катастроф. Нерівномірність розподілу цих загроз призводить до того, що найбільш вразливі верстви суспільства несуть основний тягар алгоритмічної дискримінації та маніпуляцій, тоді як бостромівський аналіз дозволяє виявити глибинну логіку відсутності контролю навіть у нинішніх системах ШІ. Ефективне управління такими динамічними викликами потребує не статичних норм, а комплексної взаємодії технічних стандартів, правового регулювання та суспільного моніторингу. Відтак, перспективи подальших досліджень мають зосередитися на розробці методологій кількісної оцінки ризиків, забезпеченні підзвітності алгоритмів та аналізі здатності демократичних інститутів встигати за темпами технологічної експансії, яка загрожує когнітивній автономії людей.

ЛІТЕРАТУРА

1. Андрощук Г. Штучний інтелект: економіка, інтелектуальна власність, загрози. Теорія і практика інтелектуальної власності. 2021. № 2. С. 56-74.
2. Бостром Нік Суперінтелект. Стратегії і небезпеки розвитку розумних машин / пер. з англ. Антон Ящук, Антоніна Ящук. Київ : Наш Формат, 2020. 408 с.
3. Колесніков А. П., Карапетян О. М. Штучний інтелект: переваги та загрози використання.// Ефективна економіка. 2023. № 8.
4. Скіцько О., Складанний П., Ширшов Р., Гуменюк М., Ворохоб М. Загрози та ризики використання штучного інтелекту // Кібербезпека: освіта, наука, техніка. 2023. № 2 (22). С. 6–18.
5. ISO 31000:2018. Risk management – Guidelines. International Organization for Standardization, 2018.

6. Khlaaf H., Mishkin P., Achiam J., Krueger G., Brundage M. A Hazard Analysis Framework for Code Synthesis Large Language Models. arXiv:2207.14157v1 [cs.SE]. 2022. 20 p.
7. Sizing the prize: What's the real value of AI for your business and how can you capitalise? PwC Global Artificial Intelligence Study. PricewaterhouseCoopers. 2017. URL: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf (дата звернення: 02.03.2026).

Фриз Артур
Науковий керівник – доц. Грушко Віктор

ШТУЧНИЙ ІНТЕЛЕКТ ЯК КАТАЛІЗАТОР НОВОЇ ХВИЛІ ГЛОБАЛІЗАЦІЇ

У статті здійснений аналіз ролі штучного інтелекту як каталізатора якісно нової хвилі глобалізації. Досліджено механізми, через які ШІ-технології трансформують глобальну торгівлю, міжнародний поділ праці та конкурентоспроможність держав. Розглянуто феномен «безмаштабної глобалізації» та її наслідки для країн глобального Півдня. Проаналізовано геополітичний вимір технологічного суперництва між США та КНР, а також суперечності між глобальним характером штучного інтелекту та прагненням держав до цифрового суверенітету. Окреслено виклики, що постають перед людством в умовах намагання контролювати глобальні процеси в той час, коли зростає небезпека можливого унезалежнення від волі людей штучного інтелекту, під впливом зростання його можливостей.

Ключові слова: штучний інтелект, глобалізація, цифровий суверенітет, технологічна нерівність, глобальне управління, «безмасштабна глобалізація».

Постановка проблеми. Глобалізація як процес поглиблення взаємозалежності між державами, ринками, суспільствами та культурами пройшла у своєму розвитку кілька якісно відмінних фаз, кожна з яких характеризувалась домінуванням певної технології. Перша хвиля, що розгорнулася у XIX столітті, спиралася на паровий двигун і залізничний транспорт. Друга хвиля, простимульована Другою світовою війною, базувалася на контейнеризації та реактивній авіації. Третя, розпочата у кінці XX ст.., поглибила глобалізацію можливостей цифрових технологій, зокрема, інтернету. Сьогодні людство стоїть на порозі нової, четвертої хвилі, рушієм якої є штучний інтелект – технологічний феномен, що за своїм трансформаційним потенціалом не має прецедентів в усій попередній історії. Питання про роль, яку здатний відіграти у житті суспільства цей фактор, його виграшні та програшні сторони, а також про ризики і загрози, що він здатний нести для глобального людства, набуває першочергового значення для існуючої цивілізації.

Аналіз досліджень та публікацій. Дана тематика одержала свій розвиток у працях Е. Бринйолфссона та Е. МакАфі, які дослідили економічні наслідки цифрової революції; Р. Болдуіна, що здійснив глибокий аналіз проблем глобалізації в умовах роботизації; М. Кастельса, який розробив теорію мережевого суспільства. Геополітичний вимір ШІ-суперництва аналізується у працях К.-Ф. Лі та у звітах McKinsey Global Institute і UNCTAD.