УДК: 004.021:004.89:004.056.5 DOI: **10.33099/2311-7249/2026-55-1-74-83**

ЖИВИЛО Євген Олександрович,
кандидат наук з державного управління, доцент,
Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна,
https://orcid.org/0000-0003-4077-7853

КУЧМА Юрій Володимирович,
кандидат технічних наук, доцент,
Товариство з обмеженою відповідальністю приватний вищий навчальний заклад
«Університет сучасних технологій», Київ, Україна,
https://orcid.org/0009-0002-5498-4271

МУЛЬТИАГЕНТНА МОДЕЛЬ АДАПТИВНОЇ ДОВІРИ В ДЕЦЕНТРАЛІЗОВАНИХ КОНФІДЕНЦІЙНИХ СИСТЕМАХ ПІД ВПЛИВОМ АТАК НА ЦІЛІСНІСТЬ ОБЧИСЛЮВАЛЬНИХ ПРОЦЕСІВ

У статті розглянуто проблему забезпечення цілісності даних у децентралізованих конфіденційних системах за умов відсутності централізованого контролю та можливості адаптивних атак. Запропоновано мультиагентну модель адаптивної довіри, що дозволяє динамічно оцінювати надійність учасників і знижувати ризики порушення цілісності обчислювальних процесів під час міжорганізаційної взаємодії.

***Метою статті** є розроблення мультиагентної моделі адаптивної довіри для децентралізованих конфіденційних систем, здатної забезпечувати цілісність та достовірність обчислювальних процесів за умов наявності адаптивних атак на вузли мережі.*

***Методи дослідження.** Під час проведення дослідження застосовано методи аналізу та синтезу для вивчення підходів до побудови мультиагентних систем і механізмів управління довірою в децентралізованих середовищах. Метод системного та імітаційного моделювання використано для розроблення мультиагентної моделі адаптивної довіри та дослідження її поведінки в умовах атак на цілісність обчислювальних процесів. Експериментальні та порівняльні методи дозволили оцінити ефективність запропонованого підходу й обґрунтувати його переваги над статичними моделями довіри.*

***Отримані результати дослідження.** В статті формалізовано атаки на цілісність обчислювальних процесів і розроблено мультиагентну модель адаптивної довіри для децентралізованих конфіденційних систем на основі байєсівського оновлення та еволюційної адаптації стратегій. Результати симуляційних експериментів підтвердили, що запропонована модель забезпечує високу стійкість до атак, швидку стабілізацію рівнів довіри агентів і ефективний баланс між безпекою, конфіденційністю та продуктивністю.*

***Елементи наукової новизни.** У роботі удосконалено підходи до формування довіри в децентралізованих системах шляхом інтеграції моделей багатоагентної взаємодії та стохастичної ігрової теорії, у межах яких довіру подано як еволюційний процес за умов неповної інформації. Розширено відомі байєсівські моделі довіри за рахунок поєднання механізмів байєсівського оновлення переконань з алгоритмами підкріплювального навчання, що забезпечує динамічну адаптацію поведінки агентів до змінних і цілеспрямованих атак на цілісність обчислювальних процесів. Уточнено механізм корекції стратегій агентів, який поширює класичні ігрові моделі довіри на децентралізовані конфіденційні системи без централізованого контролю, підвищуючи їхню стійкість до адаптивних загроз.*

***Теоретичне та практичне значення викладеного у статті.** Дослідження розширює теоретичні підходи до формування адаптивної довіри у децентралізованих системах та інтегрує байєсівське оновлення з алгоритмами підкріплювального навчання. Практично модель підвищує стійкість до атак на цілісність і забезпечує конфіденційність обміну даними, що дозволяє будувати адаптивно захищені платформи у federated learning, Web3 та IoT.*

***Ключові слова:** мультиагентна система, адаптивна довіра, децентралізована взаємодія, цілісність даних, атаки на цілісність, кіберстійкість, кібербезпека.*

Вступ

Постановка проблеми. Сучасні децентралізовані системи оброблення даних, що функціонують у режимі розподіленої взаємодії між автономними вузлами, дедалі частіше використовуються для міжорганізаційного обміну результатами обчислень, моделей машинного навчання та аналітичних рішень без передачі вихідних даних. Такий підхід забезпечує високий рівень конфіденційності, однак створює нові

виклики для збереження цілісності та достовірності обчислювальних процесів у середовищах із неоднорідним рівнем довіри. Проте низка досліджень показала, що такі системи залишаються вразливими до атак на цілісність (integrity attacks), включно з Byzantine-атаками, backdoor-впровадженням, модифікацією оновлень клієнтів та отруєнням даних (data poisoning) [1; 2; 3].

Основна проблема полягає в тому, що більшість існуючих підходів спираються на статичні або попередньо задані моделі довіри, які не враховують динамічну поведінку агентів під час атак, змін у середовищі або реакції на систему захисту. Наприклад, адаптивні зловмисники, які змінюють свою тактику залежно від заходів захисту, можуть уникати простих детекторів і алгоритмів усереднення ваг оновлень (aggregation) [4].

Ще одним викликом є гетерогенність даних (non-IID), що ускладнює ідентифікацію відхилень у поведінці агентів, оскільки "нормальна" поведінка одного вузла може виглядати як аномалія в контексті іншого через різницю в розподілах даних [5; 6].

Відсутність централізованого контролера позбавляє можливості глобального моніторингу або централізованої верифікації внесків учасників, що робить системи особливо вразливими до адаптивних атак, коли зловмисник використовує знання про захисні механізми або історію дії системи, щоб уникнути виявлення [4].

Аналіз останніх досліджень і публікацій. Сучасні децентралізовані конфіденційні системи, включно з платформами федеративного навчання, IoT-екосистемами та Web3-інфраструктурами, дедалі частіше стають об'єктом досліджень у контексті забезпечення цілісності даних та адаптивної довіри між вузлами. Ряд міжнародних дослідників, таких як G. Chen, F. Zeng, J. Zhang та інші, підкреслюють необхідність застосування адаптивних моделей довіри, здатних динамічно оцінювати надійність агентів на основі їх поведінки та взаємодії, що дозволяє суттєво знижувати ймовірність порушення цілісності в IoT-середовищах. Цей підхід акцентує увагу на необхідності інтеграції поведінкових індикаторів для формування механізмів довіри.

Водночас дослідження [7] демонструє, що у децентралізованих системах адаптивні стратегії консенсусу та самоорганізації агентів стають критично важливими для відновлення довіри після атак на цілісність. Автори наголошують на доцільності інтеграції алгоритмічних методів аналізу та керування із концепцією багатоагентних систем, що дозволяє підвищити стійкість децентралізованих структур до непередбачуваних і адаптивних дій зловмисників у динамічному інформаційному середовищі. Такий підхід дозволяє переходити від статичних моделей довіри до динамічних, що еволюціонують залежно від стану системи та історії взаємодій.

Іншим напрямом є розвиток, пов'язаний із застосуванням блокчейн-технологій та ШІ, для підвищення прозорості та верифікації транзакцій у смарт-містах та інших критичних інфраструктурах. Так, в дослідженні [8] було запропоновано децентралізовану рамку довіри, що інтегрує алгоритми AI для виявлення загроз у реальному часі, забезпечуючи баланс між конфіденційністю та цілісністю даних. Цей підхід демонструє ефективність комбінованих методів у підвищенні кіберстійкості систем із розподіленою архітектурою.

Крім академічних джерел, суттєве значення для формування методологічного підґрунтя дослідження мають рекомендації, протоколи та стандарти міжнародних інституцій, зокрема NIST, ISO/IEC, IEEE та Європейського агентства з кібербезпеки (ENISA), які визначають рамкові принципи забезпечення довіри, цілісності та адаптивності в децентралізованих обчислювальних середовищах. Так, NIST в рамках своїх звітів з кібербезпеки підкреслює важливість реалізації Zero Trust-підходів у Web3 та децентралізованих системах, акцентуючи на необхідності забезпечення автентичності і цілісності даних за відсутності централізованого контролю [9]. Подібні підходи інтегруються у практику кіберстійкості в Європейському Союзі, де Стратегія кібербезпеки ЄС та Акт про кіберстійкість визначають вимоги до управління ризиками, сертифікації та реагування на інциденти [10].

Варто відзначити, що адаптивні моделі довіри набувають особливої значущості в українських дослідженнях і практиках кіберзахисту. За результатами аналізу досліджень встановлено, що кібератаки на критичну інфраструктуру України свідчать про високий рівень операційної стійкості, досягнутий завдяки інтеграції адаптивних стратегій довіри, гнучкого управління ризиками та своєчасного реагування на загрози в рамках децентралізованих систем моніторингу [11]. Додатково, численні корпоративні та аналітичні звіти NIST демонструють практичну ефективність підходів, які комбінують алгоритмічну оцінку довіри, байєсівське оновлення ймовірностей та механізми підкріплювального навчання для адаптивного реагування на атаки [12].

В цілому проведений аналіз літературних та практичних джерел дозволяє зробити висновок, що сучасні підходи до формування довіри у децентралізованих конфіденційних системах орієнтовані на інтеграцію динамічних адаптивних механізмів, алгоритмів ШІ та криптографічних протоколів для забезпечення цілісності даних. Проте залишається низка відкритих питань, зокрема щодо оптимізації параметрів адаптації, масштабованості моделей та інтеграції українського та європейського нормативного досвіду у реальні системи. Це формує основу для подальших наукових досліджень у напряму мультиагентної моделі адаптивної довіри.

Метою статті є розроблення мультиагентної моделі адаптивної довіри для децентралізованих

конфіденційних систем, здатної забезпечувати цілісність та достовірність обчислювальних процесів за умов наявності адаптивних атак на вузли мережі.

Дослідження спрямоване на інтеграцію поведінкових, ймовірнісних та криптографічних механізмів для формування динамічної системи оцінки довіри, яка здатна коригувати поведінку агентів у режимі реального часу залежно від виявлених загроз і контексту взаємодії.

Виклад основного матеріалу дослідження

Децентралізовані конфіденційні системи, що функціонують без централізованого контролера, формують сучасну парадигму обробки та обміну даними між організаціями/установами, забезпечуючи високий рівень конфіденційності та мінімізуючи ризики витоку інформації.

Проте, паралельно з розширенням функціональних можливостей таких систем зростає і їх вразливість до адаптивних атак на цілісність, що включають "отруєння даних", підміни результатів обчислень, інжекцію фальшивих оновлень та інші форми цілеспрямованих порушень. Актуальність розробки механізмів адаптивної довіри зумовлена потребою не лише ідентифікувати потенційні загрози, але й ефективно коригувати поведінку агентів у динамічному середовищі, забезпечуючи стійкість системи до маніпуляцій без порушення конфіденційності.

Виходячи з викладеного, слід підкреслити, що різноманітність та складність атак на обчислювальні процеси зумовлюють необхідність системного підходу до їхньої класифікації та формалізації, що, у свою чергу, створює передумови для розробки ефективних методів виявлення та протидії.

Першим класом атак є "отруєння даних", яке передбачає внесення шкідливих змін у локальні набори даних агентів з метою викривлення глобального результату обчислень. Формально, для агента i локальний набір даних D_i трансформується у $\widehat{D_i} = D_i \cup \Delta_i$, де Δ_i містить некоректні або модифіковані записи. Ця модифікація призводить до зміни локальних параметрів w_i, а в глобальній функції агрегації $\mathcal{F}(\{\widehat{w_i}\})$ виникає систематичне викривлення, що порушує цілісність моделі [13]. Неперервно з цим розглядається клас маніпуляцій градієнтами (gradient manipulation/model poisoning), коли агент змінює локальні градієнти ∇w_i шляхом додавання шкідливого відхилення ε_i. У результаті глобальна модель $\widehat{w_i} = \frac{1}{N}\sum_i(\Delta w_i + \varepsilon_i)$ демонструє зміщення, що призводить до порушення цілісності обчислень без необхідності модифікувати "сирі дані". Такий тип атак підкреслює важливість оцінки поведінкової динаміки агентів у процесі адаптивної довіри [14].

У продовження цієї логіки виділяється клас підміни результатів обчислень (result tampering), коли агент змінює локально обчислений результат $r_i = f(D_i)$ перед його передачею у систему. Зловмисна модифікація $\widehat{r_i} = r_i + \delta_r$ може залишатися непоміченою в умовах відсутності централізованого верифікатора, що підкреслює критичну потребу у впровадженні механізмів контролю та оцінки на рівні взаємодіючих агентів.

Паралельно з формалізацією класів атак розглядається адаптивна поведінка агентів, яка визначається здатністю вузлів змінювати свої дії у відповідь на стан системи, історію взаємодій та реакції захисних механізмів. Формально агент i характеризується станом $s_i(t)$ та політикою дій $\pi_i(s_i(t))$, де $\pi_i: s_i(t) \to a_i(t), a_i(t) \in \{\text{коректна дія, шкідлива дія}\}$.

Адаптивні агенти модифікують свою поведінку залежно від історії $H_i = \{s_i(\tau), a_i(\tau)\}_{\tau=0}^{t-1}$, що ускладнює ідентифікацію атак стандартними детекторами та вимагає інтеграції алгоритмів самоадаптації у систему оцінки довіри [15].

Таким чином, систематизована формалізація класів атак і моделей поведінки агентів формує теоретичну основу для побудови мультиагентної моделі адаптивної довіри, яка дозволяє не лише виявляти потенційні загрози, а й динамічно коригувати поведінку вузлів для забезпечення цілісності та стійкості децентралізованої системи.

Для реалізації механізму адаптивної довіри необхідно побудувати модель, яка б ураховувала неповноту інформації щодо агентів, їхню історію поведінки та локальні перевірки, та забезпечувала еволюційну динаміку стратегій взаємодії. Така модель має спиратися на еволюційну теорію ігор з неповною інформацією, що дозволяє формалізувати змінну довіру як функцію спостережених індикаторів та обмежених знань агентів.

Отже, нехай множина агентів позначається як $\mathcal{A} = \{1, 2, \dots, N\}$. При цьому кожен агент $i \in \mathcal{A}$ зберігає історію взаємодій $H_i = \left\{\left(j, a_j(\tau), s_j(\tau)\right) \middle| \tau = 0, \dots, t-1\right\}$, де j – партнер, $a_j(\tau)$ – дія агента j (наприклад, величина градієнту, відхилення від агрегованої моделі тощо). Агент i формує свою оцінку довіри $T_i(t)$ щодо партнера j за допомогою функції довіри.

$$T_{i,j}(t) = f\left(T_{i,j}(t-1), \phi\left(a_j(t-1), s_j(t-1)\right), v_i(j,t)\right), \quad (1)$$

де ϕ – оцінка поведінкових індикаторів (наприклад, відхилення у внесках або градієнтах), а $v_i(j,t)$ – локальна перевірка надійності, яку агент i може виконати на основі агрегованої інформації або порівнянь.

Оскільки інформація про агента j неповна, функція $T_{i,j}(t-1)$ повинна бути адаптивною й побудованою в рамках моделі з неповною інформацією із

застосуванням байєсівського підходу. Зокрема, агент i має апріорні переконання $P_i(type_j)$ про тип агента j, де тип включає характеристики якості внеску, ризику маніпуляції, довіри. При кожній взаємодії агент оновлює свої переконання за правилом Байєса:

$$P_i\left(type_j \middle| H_i(t)\right) \propto P_i\left(type_j \middle| H_i(t-1)\right) \cdot L\left(a_j(t-1), s_j(t-1) \middle| type_j\right), \quad (2)$$

де L – ймовірність спостереження поведінки агента j, якщо він має тип $type_j$.

Еволюційна компонента моделі полягає в тому, що агенти приймають стратегії взаємодії $\pi_i(t)$ із урахуванням оцінок довіри, вибираючи між "чесною" та "обережною/зловмисною" моделлю дій залежно від ризику. Формально:

$$\pi_i(t) = \begin{cases} \text{визначена кооперативна дія, якщо } T_{i,j}(t) \geq \theta_i(t) \\ \text{обережна/захисна дія, якщо } T_{i,j}(t) < \theta_i(t) \end{cases}$$

де $\theta_i(t)$ – порогове значення довіри, яке може бути адаптивним і залежати від макро-стану системи, історії атак та поведінки інших агентів.

Як показано в роботах із моделювання ігор довіри за умов неповної інформації [16], асиметричні взаємодії між агентами змінюють динаміку стратегій, що покладено в основу розробленої моделі адаптивної довіри у присутності атак.

Для забезпечення стійкості та запобігання маніпуляціям модель може бути доповнена компонентами покарання або штрафу для агентів, які виявились ненадійними у перевірках довіри. Стратегія з елементами покарання сприяє еволюційному домінуванню надійних агентів у популяціях, де відбуваються періодичні взаємодії і переперевірки.

Отже, розроблена математична модель адаптивної довіри являє собою інтегровану конструкцію, що поєднує байєсівське оновлення переконань щодо типів агентів із механізмами оцінки поведінкових індикаторів і результатів локальних аудитів, на основі яких формується еволюційна динаміка стратегій взаємодії за змінних порогів довіри, доповнена компонентом покарання та винагороди, який забезпечує корекцію поведінки агентів і сприяє стабілізації системи в умовах атак на її цілісність.

У процесі функціонування децентралізованих конфіденційних систем одним із ключових чинників забезпечення стійкості до атак на цілісність є формування адаптивного механізму довіри між автономними агентами. Відсутність центрального контролю зумовлює необхідність розподіленої оцінки надійності кожного вузла, що діє в умовах неповної інформації про стан системи. У таких середовищах модель довіри повинна не лише відображати поточний рівень коректності поведінки агентів, але й динамічно адаптуватися до змін у структурі взаємодій та виявлених аномалій.

З математичної точки зору, взаємодію агентів можна описати в термінах еволюційної гри з неповною інформацією, де кожен учасник має обмежені знання щодо типів та стратегій інших агентів, а його рішення формуються на основі апостеріорної оцінки довіри. Нехай множина агентів визначається як $A = \{a_1, a_2, \dots, a_n\}$, а кожен агент a_i характеризується набором станів S_i та стратегією поведінки π_i, що оновлюється в часі t відповідно до спостережуваних результатів взаємодій. Тоді рівень довіри $T_{ij}(t)$ між агентами a_i та a_j може бути подано як функцію:

$$T_{ij}(t+1) = f\left(T_{i,j}(t), \Delta h_{ij}(t), \beta_i, \gamma_i\right), \quad (3)$$

де $\Delta h_{ij}(t)$ – зміна історії взаємодій, β_i – коефіцієнт ваги поведінкових індикаторів, а γ_i – коефіцієнт впливу локальних перевірок [17].

Динаміка еволюції стратегій у такій грі визначається за принципом реплікаторного рівняння, що описує тенденцію агентів до наслідування стратегій із вищою вигодою. Таким чином, на кожному кроці система самокоригується, зміщуючи баланс на користь надійних учасників. У цьому контексті адаптивна довіра виступає функцією як історичних даних, так і поточних метрик ризику, що оцінюються за допомогою локальних аудиторів або незалежних вузлів-спостерігачів.

Крім того, до моделі вводиться механізм соціального підкріплення – набір правил покарання та винагороди, що регулює зміну стратегій у залежності від ступеня відповідності поведінки агента очікуваним параметрам системи. Цей компонент виконує роль зворотного зв'язку, який забезпечує стійкість мережевої взаємодії до "отруєння" даних та підміни результатів обчислень [18].

У підсумку, розроблена модель адаптивної довіри дозволяє описати процес міжагентної взаємодії як когнітивно-еволюційний цикл, у межах якого довіра формується, оновлюється та підлаштовується під змінні загрози. Такий підхід забезпечує підвищення стійкості децентралізованих систем до атак на цілісність, зберігаючи автономність та конфіденційність обміну даними між учасниками без потреби у централізованій валідації.

Розвиток концепції адаптивної довіри в децентралізованих конфіденційних системах неможливий без реалізації механізмів самоадаптації та інтелектуальної оптимізації поведінки агентів, які дозволяють системі ефективно реагувати на зміни середовища та динаміку атак на цілісність даних. У цьому контексті ключову роль відіграють алгоритми, що поєднують байєсівське оновлення ймовірностей довіри з методами підкріплювального навчання (reinforcement learning), формуючи гібридну архітектуру прийняття рішень [19].

У межах запропонованого підходу кожен агент

розглядається як автономний інтелектуальний вузол, здатний оцінювати достовірність своїх партнерів через апостеріорну оцінку довіри, що базується на байєсівській формулі оновлення переконань:

$$P(H_i|D_t) = \frac{P(D_t|H_i)P(H_i)}{\sum_j P(D_t|H_j)P(H_j)}, \quad (4)$$

де $P(H_i|D_t)$ – оновлена ймовірність надійності агента H_i після спостереження даних D_t, а $P(D_t|H_i)$ – ймовірність отримання відповідних даних за умови гіпотези про довірену поведінку агента. Такий механізм дозволяє агентам поступово уточнювати рівень довіри до своїх сусідів, формуючи індивідуальні профілі надійності, які виступають основою для подальшої оптимізації дій у системі.

Паралельно із байєсівським оновленням застосовується підкріплювальне навчання, яке забезпечує адаптацію агентів до мінливого середовища через механізм зворотного зв'язку. Кожен агент, виконуючи дію a_t у стані s_t, отримує винагороду r_t, що сигналізує про якість прийнятого рішення. Процес оновлення політики відбувається за правилом Q-навчання:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right], \quad (5)$$

де α – коефіцієнт навчання, а γ – параметр дисконтування, що визначає значущість майбутніх винагород. Таким чином, агент формує оптимальну політику π^*, яка мінімізує ризики взаємодії з потенційно шкідливими вузлами.

З метою підвищення стійкості до адаптивних атак, у систему впроваджується механізм мета-навчання, який дозволяє агентам швидко перебудовувати політику у відповідь на зміну патернів поведінки противника. Цей підхід відповідає сучасним дослідженням у сфері multi-agent reinforcement learning, де адаптація відбувається на рівні колективної оптимізації стратегій у просторі довіри.

Таким чином, впроваджені алгоритми самоадаптації забезпечують когнітивну еволюцію системи довіри, у якій агенти не лише реагують на зміни середовища, але й прогнозують потенційні аномалії, використовуючи ймовірнісні та навчальні механізми. Це дозволяє підтримувати баланс між швидкістю реакції, точністю оцінок і обчислювальною ефективністю, що є критично важливим у контексті децентралізованих безконфіденційних систем, орієнтованих на міжорганізаційну взаємодію без централізованого обміну даними.

В умовах децентралізованого середовища, де агенти функціонують без централізованого контролю, ключовим аспектом формування адаптивної довіри є механізм винагороди, який визначає, наскільки дії агента відповідають глобальним цілям системи. Оскільки поведінка агентів формується на основі індивідуального досвіду та обмеженого знання про інших учасників, функція винагороди має відображати як ймовірнісну оцінку довіри, так і ризик компрометації, враховуючи обчислювальні, комунікаційні та приватні витрати. Саме така багатокомпонентна структура дозволяє реалізувати принцип когнітивної самоорганізації, у межах якого кожен агент оптимізує свою політику на основі локального спостереження та глобальних цілей системи.

З математичної точки зору, процес прийняття рішень агентом можна представити як функцію винагороди $r_{i,j}(t)$, що визначає доцільність взаємодії між агентом a_i та його партнером a_j. Для кожної взаємодії враховується п'ять ключових показників: довіра $T_{i,j}(t)$, ризик $R_{i,j}(t)$, корисність $U_{i,j}(t)$, витрати $C_{i,j}(t)$ та втрати приватності $\Pi_{i,j}(t)$. На основі цих параметрів узагальнена функція винагороди набуває вигляду:

$$r_{i,j}(t) = \alpha \mathcal{B}\left(U_{i,j}(t), T_{i,j}(t)\right) - \beta \mathcal{P}\left(R_{i,j}(t), T_{i,j}(t)\right) - \gamma C_{i,j}(t) - \delta \Pi_{i,j}(t), \quad (6)$$

де $\alpha, \beta, \gamma, \delta$ – вагові коефіцієнти, що визначають баланс між корисністю, ризиком, витратами та приватністю. Такий підхід дозволяє відобразити багатофакторну природу довіри та забезпечити її гнучке налаштування залежно від контексту взаємодії.

Компонента корисності $\mathcal{B}(U,T) = T^k U$ відображає залежність отриманої вигоди від рівня довіри, що дозволяє системі надавати пріоритет взаємодії з перевіреними агентами. Параметр k визначає чутливість моделі до довіри, при $k > 1$ система різкіше реагує на незначне зниження T, зменшуючи винагороду для агентів із підозрілою поведінкою. У свою чергу, компонента ризику $\mathcal{P}(R,T) = \frac{R}{T+\varepsilon}$ посилює покарання за високу ймовірність компрометації, особливо коли рівень довіри є низьким. Такий підхід створює нелінійну залежність між оцінкою ризику та довірою, що підвищує точність реагування системи на аномальні дії.

Загальна миттєва винагорода агента $r_i(t)$ визначається як сума локальних винагород від усіх взаємодій, зважених за їхньою важливістю:

$$r_i(t) = \sum_{j \in \mathcal{N}_i(t)} w_{i,j}(t) r_{i,j}(t), \quad (7)$$

де $w_{i,j}(t)$ – коефіцієнти ваги, що відображають пріоритетність або частоту взаємодії з певними партнерами. Таким чином, агент динамічно оновлює власну політику поведінки на основі накопиченого досвіду, підсилюючи співпрацю з надійними вузлами та зменшуючи взаємодію з потенційно зловмисними.

Важливо зазначити, що наведена функція

винагороди володіє низкою формальних властивостей, які забезпечують стабільність процесу навчання. По-перше, її значення є обмеженими та нормованими, що запобігає чисельній нестійкості під час ітерацій підкріплювального навчання. По-друге, залежність від змінної T робить функцію адаптивною, тобто вона миттєво реагує на зміну довіри внаслідок нових подій або атак. По-третє, гладкість компонент $\mathcal{B}$ та $\mathcal{P}$ дозволяє застосовувати градієнтні методи оптимізації в контексті алгоритмів policy-gradient та actor-critic.

У рамках алгоритмів підкріплювального навчання (Q-learning, Deep Q-Network, MARL) функція винагороди використовується для оновлення політики агента. Наприклад, у класичній формі Q-оновлення:

$$Q_i(s_t, a_t) \leftarrow Q_i(s_t, a_t) + \alpha \left[r_i(t) + \gamma \max_a Q_i(s_{t+1}, a) - Q_i(s_t, a_t) \right], \quad (8)$$

де α – коефіцієнт навчання, γ – параметр дисконтування. Таким чином, кожен агент поступово формує оптимальну стратегію взаємодії, яка мінімізує ризик співпраці з недобросовісними вузлами та максимізує глобальну стійкість системи.

Запропонована функція винагороди забезпечує баланс між довірою, ризиком, ефективністю та конфіденційністю, що робить її універсальною для використання в різних класах децентралізованих систем – від федеративного навчання до міжорганізаційних IoT-мереж. Такий підхід не лише підвищує ефективність прийняття рішень у середовищах із частковою інформацією, але й забезпечує адаптивну стійкість системи до атак на цілісність**,** формуючи основу для побудови еволюційно стабільних моделей довіри нового покоління [17].

Оцінювання ефективності запропонованої моделі адаптивної довіри потребує комплексного підходу, який поєднує математичне моделювання, статистичний аналіз і комп'ютерну симуляцію поведінки агентів у середовищі з неповною інформацією. Основною метою експериментальної верифікації є визначення рівня стійкості системи до атак, спрямованих на цілісність даних і маніпулювання моделлю довіри, а також виявлення компромісів між параметрами безпеки, ефективності та приватності. Такий підхід відповідає рекомендаціям NIST щодо тестування безпечних децентралізованих систем ШІ [19] і узгоджується з європейськими підходами ENISA до оцінювання AI Integrity Resilience у критичних інфраструктурах.

В рамках проведених симуляційних експериментів було створено агентно-орієнтоване середовище, яке забезпечує моделювання сценаріїв взаємодії між раціональними агентами та зловмисними агентами, дозволяючи досліджувати динаміку поведінки системи та оцінювати ефективність стратегій протидії загрозам. Кожен агент характеризується набором параметрів: коефіцієнтом довіри $T_i(t)$, очікуваною корисністю $U_i(t)$, ймовірністю компрометації $R_i(t)$ та коефіцієнтом адаптації $\lambda_i(t)$. Початкові значення цих параметрів ініціалізуються випадковим чином відповідно до нормального розподілу, що дозволяє оцінити поведінку системи в умовах стохастичної невизначеності. Модель реалізована як еволюційна гра з неповною інформацією, у межах якої агенти ітеративно оновлюють свої стратегії відповідно до алгоритму байєсівського підкріплення.

Ключовими показниками для кількісної оцінки ефективності моделі виступають метрики стійкості, що характеризують здатність системи зберігати функціональну цілісність у разі атак або дестабілізуючих впливів, а саме:

Integrity Resilience Index (IRI) – інтегральний показник здатності системи відновлювати достовірність обчислень після атак на дані чи моделі. Формально визначається як відношення середнього рівня цілісності системи після атаки до рівня цілісності до атаки:

$$IRI = \frac{\bar{I}_{post}}{\bar{I}_{pre}} \quad (9)$$

Значення $IRI \approx 1$ свідчить про високу толерантність системи до порушень.

Adaptive Robustness Index (ARI) – характеристика швидкості та стабільності адаптації системи до змінних умов середовища. Обчислюється як середнє нормалізоване значення похідної функції довіри за часом:

$$ARI = \frac{1}{N} \sum_{i=1}^{N} \frac{|\Delta T_i(t)|}{\Delta t} \quad (10)$$

Чим менше значення ARI, тим більш стабільною є система за наявності динамічних атак.

Privacy-Integrity Trade-off (PIT) – метрика, що характеризує компроміс між збереженням конфіденційності агентів і забезпеченням цілісності обчислень. Визначається як співвідношення між приростом довіри та втратами приватності:

$$PIT = \frac{\Delta T}{\Delta \Pi} \quad (11)$$

Високі значення PIT свідчать про ефективну роботу моделі, коли мінімальні втрати приватності забезпечують суттєве підвищення довіри в системі.

На основі серії симуляцій із варіюванням частки зловмисних агентів у межах від 5 % до 40 % встановлено, що запропонована модель демонструє високі показники стійкості як при низькому $IRI > 0.93$, так і при помірному рівні загроз $ARI < 0.15$, забезпечуючи ефективну взаємодію між агентами та збереженням стабільності системи.

На Рисунку 1 представлений графік динаміки метрик стійкості системи адаптивної довіри за різної частки зловмисних агентів.

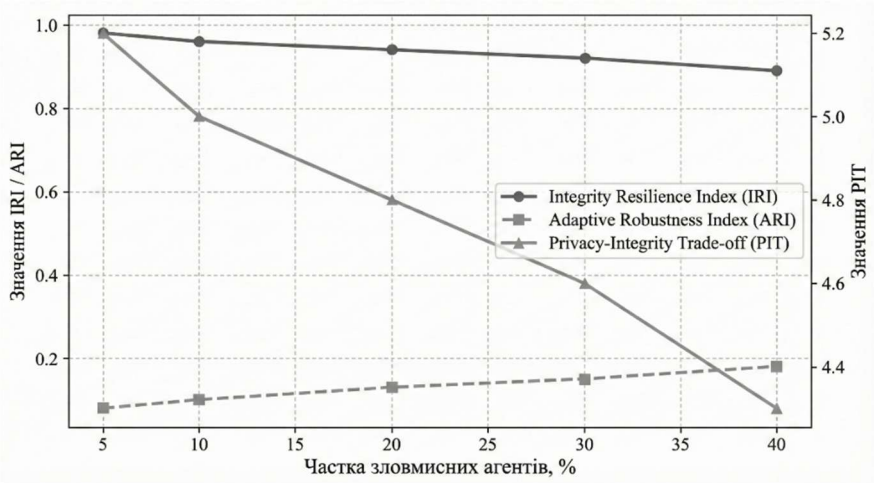

Рисунок 1 – Графік динаміки метрик стійкості системи адаптивної довіри

Отримані результати підтверджують її здатність до самовідновлення після атак типу gradient manipulation та data poisoning. При цьому середнє значення $PIT = 4.7$ свідчить про оптимальний баланс між прозорістю обміну інформацією і збереженням приватних даних агентів.

Необхідно зазначити, що подібна динаміка узгоджується з результатами досліджень Європейського інституту кібербезпеки, які відзначають ефективність байєсівських моделей довіри в децентралізованих системах за умов неповної інформації.

Як показано на Рисунку 2, у моделі з реалізованими механізмами адаптивної винагороди та покарання крива еволюції параметра $T_i(t)$ демонструє більш швидку стабілізацію, орієнтовно на 25-30% швидше, ніж у неадаптивній системі. Це свідчить про вищу когнітивну узгодженість агентів та ефективнішу компенсацію флуктуацій довіри. Амплітуда коливань у зоні стабілізації також є нижчою, що корелює зі зменшенням ентропії системи та вказує на структурну впорядкованість процесів формування довіри [20].

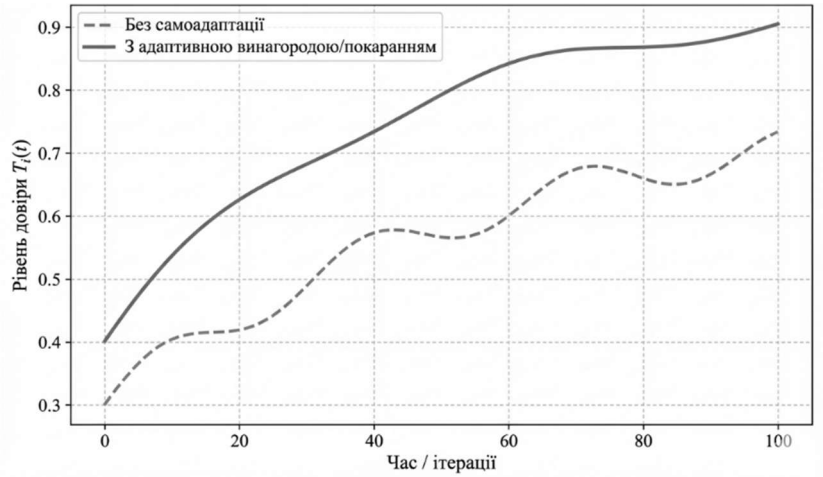

Рисунок 2 – Еволюція параметра довіри $T_i(t)$ у різних моделях агентної взаємодії

Таким чином, результати симуляційних експериментів підтверджують, що розроблена модель адаптивної довіри з байєсівським оновленням і підкріплювальним навчанням забезпечує високу ступінь стійкості до атак на цілісність, демонструє стабільну поведінку в умовах динамічних загроз і дозволяє ефективно балансувати між вимогами безпеки, конфіденційності та продуктивності. Отримані значення метрик IRI, ARI та PIT можуть бути використані як уніфіковані індикатори кіберстійкості децентралізованих систем нового покоління, що узгоджується з методологією оцінки надійності відповідно до рекомендацій NIST та ENISA.

Висновки

Проведене дослідження засвідчило, що запропонована концепція адаптивної довіри у децентралізованих обчислювальних системах є ефективним підходом до підвищення їхньої стійкості та цілісності за умов наявності зловмисних агентів і неповної інформації. Формалізація класів атак, орієнтована на маніпуляцію даними, градієнтами та результатами обчислень, дозволила структурувати потенційні загрози та закласти основу для розробки математичних імітаційних моделей поведінки агентів. Це забезпечило системне бачення динаміки довіри в середовищах із високим рівнем невизначеності.

Розроблена математична модель адаптивної довіри, побудована на засадах еволюційної гри з неповною інформацією, інтегрує байєсівське оновлення переконань, поведінкові індикатори, локальні аудити та механізми динамічного коригування стратегій агентів. Її впровадження дозволяє реалізувати самоадаптацію системи, при якій рівень довіри формується як функція історичних даних і контекстної оцінки ризику, а не лише статичних правил чи попередніх станів мережі.

Симуляційні експерименти довели, що використання підкріплювального навчання та байєсівського оновлення ймовірностей довіри суттєво підвищує стійкість системи до атак типу gradient manipulation і data poisoning. Зокрема, спостерігалось зростання показників Integrity Resilience і Adaptive Robustness, що свідчить про здатність моделі до самовідновлення після локальних порушень. Водночас дотримано оптимальне співвідношення Privacy-Integrity Trade-off, що гарантує баланс між збереженням конфіденційності агентів і забезпеченням прозорості міжвузлової взаємодії.

Результати дослідження узгоджуються з аналітичними звітами NIST (2023) та ECSO (2024), які підкреслюють доцільність використання адаптивних моделей довіри в розподілених системах без централізованого контролю. Запропонований підхід також відкриває можливості для інтеграції в Zero Trust Architecture, що набуває дедалі більшого значення в контексті корпоративної безпеки та управління критичними цифровими інфраструктурами.

Перспективи і напрями подальших досліджень полягають у розширенні моделі за рахунок мультиагентної оптимізації з урахуванням когнітивних факторів і контекстної семантики поведінки агентів, а також у застосуванні нейросимвольних методів для інтерпретації рішень у системах з неповною або недостовірною інформацією. Додаткового дослідження потребує питання метризації довіри у гібридних середовищах, де взаємодіють як людські, так і машинні агенти, з урахуванням етичних та правових аспектів прийняття рішень.

Таким чином, розроблена модель може стати базисом для побудови інтелектуальних систем довіри нового покоління, здатних не лише ідентифікувати загрози, а й еволюційно адаптуватися до змін середовища, підтримуючи цілісність, безпеку та сталий розвиток децентралізованих інформаційних інфраструктур.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Фінансування. Фінансування дослідження не здійснювалося.

Доступність даних. Дослідження виконано з використанням виключно відкритих даних, доступних у публічних джерелах

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список бібліографічних посилань

1. Sikandar H. S., Waheed H., Tahir S., Malik S. U. R., Rafique W. A Detailed Survey on Federated Learning Attacks and Defenses. *Electronics.* 2023. Vol. 12. № 2. P. 260. DOI: https://doi.org/10.3390/electronics12020260. **2. Jimenez-Gutierrez D. M., Falkouskaya Ye., Hernandez-Ramos J. L., Anagnostopoulos A., Chatzigiannakis Io., Vitaletti A.** On the Security and Privacy of Federated Learning: A Survey with Attacks, Defences, Frameworks, Applications, and Future Directions, 2025. DOI: https://doi.org/10.48550/arXiv.2508.13730. **3. Liu P., Xu X., Wang W.** Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity.* 2022. Vol. 5. № 4. DOI: https://doi.org/10.1186/s42400-021-00105-6. **4. Szeląg Ja. K., Chin Ji-J., Yip S.-Ch**. Adaptive Adversaries in Byzantine-Robust Federated Learning: A survey, Cryptology ePrint Archive, 2025. 510 p. URL: https://eprint.iacr.org/2025/510 (accessed: 10 March 2026). **5. Jimenez-Gutierrez D. M., Falkouskaya Y., Hernandez-Ramos J. L., Anagnostopoulos A., Chatzigiannakis I., Vitaletti A.** On the Security and Privacy of Federated Learning: A Survey with Attacks, Defenses, Frameworks, Applications, and Future Directions. 2025. DOI: https://doi.org/10.48550/arXiv.2508.13730 (accessed: 13 December 2025). **6. Nuria R.-B., Jiménez-López D., Luzón M. V., Herrera F., Martínez-Cámara Eu.** Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*. 2023. Vol. 90. P. 148-173, URL: https://www.sciencedirect.com/science/article/pii/S1566253522001439 (accessed: 13 December 2025). **7. Mlika F., Karoui W., Romdhane L. B.** Refined consensus mechanisms for rebuilding trust in decentralised social networks with PBFT. Expert Systems with Applications. 2025. URL: https://www.sciencedirect.com/science/article/pii/S0957417425009558 (accessed: 13 December 2025). **8. Islam R., Bose R., Roy S. et al.** Decentralized trust framework for smart cities: a blockchain-enabled cybersecurity and data integrity model. *Sci Rep.* № 15, 23454. 2025. DOI: https://doi.org/10.1038/s41598-025-06405-y. **9. Yaga D, Mell P.M.** A Security Perspective on the Web3 Paradigm. (National Institute of Standards and Technology, Gaithersburg, MD). *NIST Interagency or Internal Report (IR) NIST IR 8475*. 2025. DOI: https://doi.org/10.6028/NIST.IR.8475. **10. EU Cybersecurity Strategy, European Union**. 2025. URL: https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy (accessed: 13 December 2025). **11. Kott A., Dubynskyi G. Y., Paziuk A., Galaitsi S. E., Trump B. D., Linkov I.** Russian Cyber Onslaught Was Blunted by Ukrainian Cyber Resilience, Not Merely Security. *Computer*. 2024. Vol. 57. № 8. P. 82–89. DOI: https://doi.org/10.1109/MC.2024.3404568. **12. Cawthra J., Ekstrom M., Lusty L., Sexton Ju., Sweetnam Jo.** NIST SPECIAL PUBLICATION 1800-25, Data Integrity: Identifying and Protecting Assets. Against Ransomware and Other Destructive Events, 2020. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-25.pdf (accessed: 15 December 2025).

13. Wassim (Wes) B., El-Mahdi El-M., Usunier N. Inverting Gradient Attacks Makes Powerful Data Poisoning, 2024. URL: https://arxiv.org/pdf/2410.21453v2 (accessed: 07 December 2025). **14. Wassim (Wes) B., El-Mahdi El-M., Usunier N.** Inverting Gradient Attacks Naturally Makes Data Poisons: An Availability Attack on Neural Networks. 2024. URL: https://arxiv.org/html/2410.21453v1 (accessed: 07 December 2025). **15. Wang N., Wei D.** An Adaptive Dempster-Shafer Theory of Evidence Based Trust Model in Multiagent Systems. *Applied Sciences*. 2022. Vol. 12. № 15. 7633. DOI: https://doi.org/10.3390/app12157633. **16. Lim I. S., Masuda N.** To trust or not to trust: Evolutionary dynamics of an asymmetric N-player trust game. *IEEE Transactions on Evolutionary Computation*, 2023. Vol. 28 № 1. P. 117–131. URL: https://arxiv.org/abs/2305.01413 (accessed: 13 December 2025). **17. Kharchenko V., Fesenko H., Illiashenko O.** Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic & Computer Systems*. 2022. Vol. 2. P. 1–14. DOI: https://doi.org/10.32620/reks.2022.2.11. **18. Wang Ju., Liu Zh., Xu Yan, Li X.** (2025) Dynamic evolution in multi-player networked trust games with graded punishment. *Chaos*. 2025. Vol. 35 № 3: 033105. DOI: https://doi.org/10.1063/5.0256342. **19. National Institute of Standards and Technology (NIST).** Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD. 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1. **20. Li T., Zhang Y.** Adaptive Trust Evaluation Model Based on Entropy Weight Method for Sensing Terminal Process. *Entropy*. 2025. Vol. 27. № 2. P. 200. DOI: https://doi.org/10.3390/e27020200.

MULTI-AGENT MODEL OF ADAPTIVE TRUST IN DECENTRALISED CONFIDENTIAL SYSTEMS UNDER THE INFLUENCE OF INTEGRITY ATTACKS

ZHYVYLO Yevhen**,** Candidate of Sciences in Public Administration, Associate Professor, National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine, https://orcid.org/0000-0003-4077-7853

KUCHMA Yurii**,** Candidate of Technical Sciences, Associate Professor, Limited Liability Company Private Higher Education Institution «University of Modern Technologies», Kyiv, Ukraine, https://orcid.org/0009-0002-5498-4271

Formulation of the problem in general*. The purpose of the article is to develop a multi-agent model of adaptive trust for decentralised confidential systems, capable of ensuring the integrity and reliability of computing processes in the presence of adaptive attacks on network nodes.*

Research methods. *During the research, analysis and synthesis methods were used to study approaches to the construction of multi-agent systems and trust management mechanisms in decentralised environments. The method of system and simulation modelling was used to develop a multi-agent model of adaptive trust and to study its behaviour under attacks on the integrity of computing processes. Experimental and comparative methods enabled evaluation of the proposed approach's effectiveness and justification of its advantages over static trust models.*

Literature review. *Literary analysis shows that modern models of trust in decentralised systems are based on the integration of dynamic adaptive mechanisms, AI algorithms, and cryptographic protocols, which allow for increased cyber resilience and data integrity. At the same time, questions remain open about the scalability of models, the optimisation of adaptation parameters, and the integration of national and European regulatory approaches into practical systems, which provide a scientific perspective for the development of multi-agent models of adaptive trust.*

Research results. *The article formalises attacks on the integrity of computing processes and develops a multi-agent model of adaptive trust for decentralised confidential systems based on Bayesian updating and evolutionary adaptation of strategies. The results of the simulation experiments confirmed that the proposed model provides high resistance to attacks, rapid stabilisation of agent confidence levels and an effective balance between security, privacy and performance.*

Research novelty*. The work improves approaches to trust formation in decentralised systems by integrating models of multi-agent interaction and stochastic game theory, in which trust is modelled as an evolutionary process under conditions of incomplete information. Well-known Bayesian models of trust have been expanded by combining Bayesian belief update mechanisms with reinforcement learning algorithms, ensuring dynamic adaptation of agent behaviour to variable and targeted attacks on the integrity of computational processes. The mechanism for correcting agents' strategies has been clarified, extending classic game models of trust to decentralised, confidential systems without centralised control, thereby increasing their resistance to adaptive threats.*

Theoretical and practical significance. *The study expands theoretical approaches to the formation of adaptive trust in decentralised systems and integrates Bayesian updating with reinforcement learning algorithms. In practice, the model increases resistance to integrity attacks and ensures the confidentiality of data exchange, enabling the adaptive development of secure platforms for federated learning, Web3, and IoT.*

Conclusion and future work. *The proposed model of adaptive trust in decentralised systems, integrating Bayesian updating, behavioural indicators, and reinforcement learning, ensures agent self-adaptation and increases resistance to attacks on data integrity under conditions of incomplete information. Simulation experiments confirmed the model's effectiveness in balancing security, privacy, and the transparency of interaction, opening the way for integration into Zero Trust Architecture and the development of intelligent, next-generation trust systems.*

Keywords**:** *multi-agent system, adaptive trust, decentralised interaction, data integrity, integrity attacks, cyber resilience, cybersecurity.*

References

1. Sikandar, H. S., Waheed, H., Tahir, S., Malik, S. U. R., Rafique, W., (2023). A Detailed Survey on Federated Learning Attacks and Defenses. *Electronics 2023*. 12(2), 260. DOI: https://doi.org/10.3390/electronics12020260. **2. Jimenez-Gutierrez, D. M., Falkouskaya, Ye., Hernandez-Ramos, J. L., Anagnostopoulos, A., Chatzigiannakis, Io., Vitaletti, A.,** (2025). *On the Security and Privacy of Federated Learning: A Survey with Attacks, Defenses, Frameworks, Applications, and Future Directions*. DOI: https://doi.org/10.48550/arXiv.2508.13730. **3. Liu, P., Xu, X. & Wang, W.,** (2022). Threats, attacks and defences to federated learning: issues, taxonomy and perspectives. *Cybersecurity 5,* 4. DOI: https://doi.org/10.1186/s42400-021-00105-6. **4. Szeląg, Ja. K., Chin, Ji-J., Yi,p S.-Ch.,** (2025). Adaptive Adversaries in Byzantine-Robust Federated Learning: A survey, Cryptology ePrint Archive, 510. [online]. Available at: https://eprint.iacr.org/2025/510 [Accessed: 10 March 2026]. **5. Jimenez-Gutierrez D. M., Falkouskaya Y., Hernandez-Ramos J. L., Anagnostopoulos A., Chatzigiannakis I., Vitaletti A..** (2025). *On the Security and Privacy of Federated Learning: A Survey with Attacks, Defences, Frameworks, Applications, and Future Directions* DOI: https://doi.org/10.48550/arXiv.2508.13730. **6. Nuria, R.-B., Jiménez-López, D., Luzón, M. V., Herrera, F., Martínez-Cámara, Eu.,** (2023). Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges [online]. *Information Fusion*. 90, 148-173. Available at: https://www.sciencedirect.com/science/article/pii/S1566253522001439 [Accessed: 13 December 2025]. **7. Mlika, F., Karoui, W., Romdhane, L. B.,** (2025). Refined consensus mechanisms for rebuilding trust in decentralized social networks with PBFT [online]. *Expert Systems with Applications*. Vol. 280, Available at: https://www.sciencedirect.com/science/article/pii/S0957417425009558 [Accessed: 13 December 2025]. **8. Islam, R., Bose, R., Roy, S. et al.,** (2025). Decentralized trust framework for smart cities: a blockchain-enabled cybersecurity and data integrity model. *Sci Rep* 15, 23454. DOI: https://doi.org/10.1038/s41598-025-06405-y. **9. Yaga, D, Mell, P.M.,** (2025). A Security Perspective on the Web3 Paradigm. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) NIST IR 8475. DOI: https://doi.org/10.6028/NIST.IR.8475. **10. EU Cybersecurity Strategy, European Union**, (2025) [online]. Available at: https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy [Accessed: 13 December 2025]. **11. Kott, A., Dubynskyi, G. Y., Paziuk, A., Galaitsi, S. E., Trump, B. D. and Linkov, I.,** (2024). Russian Cyber Onslaught Was Blunted by Ukrainian Cyber Resilience, Not Merely Security. *Computer*. 57, 8, 82-89. DOI: https://doi.org/10.1109/MC.2024.3404568. **12. Cawthra, J., Ekstrom, M., Lusty, L., Sexton, Ju., Sweetnam, Jo.,** (2020). NIST SPECIAL PUBLICATION 1800-25, Data Integrity: Identifying and Protecting Assets. Against Ransomware and Other Destructive Events [online]. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-25.pdf [Accessed: 15 December 2025]. **13. Wassim, (Wes) B., El-Mahdi, El-M., Usunier, N.,** (2024). *Inverting Gradient Attacks Makes Powerful Data Poisoning* [online]. Available at: https://arxiv.org/pdf/2410.21453v2 [Accessed: 7 December 2025]. **14. Wassim (Wes) B., El-Mahdi El-M., Usunier N.,** (2024). *Inverting Gradient Attacks Naturally Makes Data Poisons: An Availability Attack on Neural Networks* [online]. Available at: https://arxiv.org/html/2410.21453v1 [Accessed: 7 December 2025]. **15. Wang, N., & Wei, D.,** (2022). An Adaptive Dempster-Shafer Theory of Evidence Based Trust Model in Multiagent Systems. *Applied Sciences*, 12(15), 7633. DOI: https://doi.org/10.3390/app12157633. **16. Lim, I. S., & Masuda, N.,** (2023). To trust or not to trust: Evolutionary dynamics of an asymmetric N-player trust game [online]. *IEEE Transactions on Evolutionary Computation*, 28(1), 117-131. Available at: https://arxiv.org/abs/2305.01413 [Accessed: 13 December 2025]. **17. Kharchenko, V., Fesenko, H. & Illiashenko, O.,** (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic & Computer Systems*, 2, P. 1-14. DOI: https://doi.org/10.32620/reks.2022.2.11. **18. Wang, Ju., Liu, Zh., Xu, Yan, Li, X.,** (2025). Dynamic evolution in multi-player networked trust games with graded punishment. *Chaos,* 35(3): 033105. DOI: https://doi.org/10.1063/5.0256342. **19. National Institute of Standards and Technology (NIST).** (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD. DOI: https://doi.org/10.6028/NIST.AI.100-1. **20. Li, T., & Zhang, Y.,** (2025). Adaptive Trust Evaluation Model Based on Entropy Weight Method for Sensing Terminal Process. *Entropy,* 27(2), 200. DOI: https://doi.org/10.3390/e27020200.

Рукопис надійшов до редакції *21.01.2026*
Рукопис прийнято до друку після рецензування *27.03.2026*
Дата публікації *30.04.2026*