«Системні технології» 5 (166) 2026 Том 2 «System technologies»
DOI 10.34185/1562-9945-5-166-2026-28
УДК: 004.056.5

С.Б. Бурченко

ЕКСПЕРИМЕНТАЛЬНА АГЕНТНА СИСТЕМА КІБЕРЗАХИСТУ НА ЦИФРОВОМУ ДВІЙНИКУ ЕНЕРГЕТИЧНОГО ОБ'ЄКТА ТА КІЛЬКІСНИЙ МЕТОД ВИЗНАЧЕННЯ ДОПУСТИМОГО РІВНЯ АВТОНОМНОСТІ

Анотація. Посилення взаємозв'язку між інформаційними мережами та системами управління на енергетичних об'єктах підвищує вимоги до швидкості виявлення кіберзагроз, водночас створюючи ризик того, що автоматизовані заходи реагування можуть негативно вплинути на технологічний процес. У статті запропоновано концептуальну архітектуру агентної системи кіберзахисту, що функціонує на базі цифрового двійника енергетичного об'єкта, а також кількісний метод визначення допустимого рівня автономності захисних дій. Запропонована архітектура об'єднує підсистеми збору та нормалізації телеметрії, граф активів, агентів моніторингу й планування, а також детермінований механізм перевірки та виконання політик безпеки. Цифровий двійник слугує контрольованим ізольованим середовищем для моделювання станів системи та сценаріїв атак, а також для валідації рішень перед їх застосуванням у реальній (операційній) мережі. Метод враховує критичність активів, масштаб потенційних кіберфізичних наслідків, граничні значення похибок (на етапах виявлення, атрибуції, планування та виконання), якість телеметрії, а також розбіжність між цифровим двійником і фізичною системою. Для дозволених дій розраховується безрозмірний «індекс ризикового тиску», на основі якого визначається рівень автономності — від A0 до A4. Детерміновані обмеження безпеки мають пріоритет: команди, заборонені політикою або відсутні у «білому списку», блокуються незалежно від значення індексу; крім того, автономне виконання дій забороняється у разі критичної втрати телеметрії або виявлення неідентифікованої атаки. Система передбачає підтвердження дій оператором, суворе розмежування операційного та тестового середовищ, використання гістерезису та ведення журналу аудиту. Наукова новизна підходу полягає в поєднанні формалізованих правил делегування дій агенту з перевіркою обмежень безпеки та врахуванням похибок цифрового двійника — у межах єдиної системи політик. Наведено ілюстративні сценарії та план подальшої валідації, що включає порівняння з ручним реагуванням і системами, які базуються на жорстких правилах, а також перевірку можливості скасування безпечних дій та оцінювання невизначеності. Стаття має концептуально-методологічний характер: наведені параметри та порогові значення є ілюстративними; експериментальні результати не розглядаються. Питання практичної доцільності, безпеки та впливу на час реагування потребують уточнення після впровадження цифрового двійника та проведення незалежного тестування в ізольованому середовищі.

© Бурченко С.Б., 2026
Copyright for this paper by its authors. Use permitted under License CC BY 4.0.
ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Ключові слова: Agentic AI, цифровий двійник, кібербезпека, OT security, енергетика, автономне реагування, кіберфізичний ризик, IEC 62443.

1. Вступ. Енергетична інфраструктура дедалі частіше об'єднує корпоративні IT мережі, диспетчерські системи та польові контролери. У такому середовищі швидке виявлення аномалій є необхідним, але автономна дія штучного інтелекту може сама створити небезпечний технологічний стан. Тому практичне питання полягає не лише у точності моделі, а й у тому, який клас дій вона має право виконувати щодо конкретного активу [1–3, 17, 18].

Мета роботи — запропонувати архітектуру агентного кіберзахисту й відтворюване правило обмеження автономності за критичністю активу, помилками агентного контуру, наслідками та похибкою цифрового двійника. Стаття є методологічною пропозицією: експеримент і валідацію ще не проведено; числові сценарії є лише умовними прикладами [7, 11, 12].

2. Пов'язані стандарти та обмеження. Архітектуру узгоджено з принципом сегментації зон і каналів IEC 62443 та з підходом NIST до виявлення, реагування і відновлення. Модель не отримує прямого привілейованого каналу до пристроїв безпеки. Захисні SIS, аварійне відключення, релейний захист і будь-які команди, що змінюють уставки або стан силового обладнання, виключено з автономного контуру. [1–3, 7, 10]

Передбачено використання цифрового двійника як ізольованого середовища для перевірки політик. Планується моделювати топологію, телеметрію та зв'язки між кіберподіями і технологічними станами; відповідність моделі фізичному процесу потребує валідації [11, 17].

3. Огляд сучасних досліджень і наукова прогалина. Огляди 2024–2025 років підтверджують, що цифрові двійники дедалі частіше використовують як середовище для моделювання атак, прогнозування стану та перевірки захисних політик [11, 12]. Для енергетики це поєднання особливо важливе: кіберінцидент може впливати не лише на конфіденційність або доступність даних, а й на безпечність технологічного процесу, безперервність електропостачання та ресурс обладнання [17, 18].

Паралельно формується напрям agentic AI для операційної підтримки захисту. Праці з критичної інфраструктури наголошують на потенціалі LLM-агентів для кореляції доказів, автоматизованої аналітики та оцінювання сценаріїв [12]. Однак нові огляди загроз до AI-агентів демонструють, що автономний контур сам розширює площину атаки: prompt injection через неочищені журнали, компрометація інструментів, tool abuse, витік секретів, отруєння контексту та помилкова взаємодія агентів [13, 19, 20].

Роботи з human-in-the-loop навчання та пояснюваного AI у промислових кіберфізичних системах підтримують принцип делегування дій за умов перевірюваності, зрозумілого обґрунтування й можливості скасування [14, 15]. Окремі дослідження часових аномалій створюють методичну основу для оцінювання помилки детекції у технологічній телеметрії [16]. Разом ці праці пояснюють, чому точність детектора не може бути єдиною підставою для автономного реагування.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Аналіз використаних джерел виявляє потребу в подальшому дослідженні правила делегування дій, яке одночасно враховує критичність активу, кіберфізичний наслідок, помилки агентного контуру та розбіжність цифрового двійника зі стендом. Запропонований індекс є дослідницькою гіпотезою; повну новизну відносно всієї літератури необхідно підтвердити систематичним оглядом.

4. Експериментальна архітектура. Запропонована архітектура складається з п’яти логічних шарів: збирання телеметрії, графа активів, агента спостереження, агента планування та виконавця політики. Планується використовувати журнали HMI й інженерної станції, NetFlow, події автентифікації, ознаки OPC UA та Modbus TCP і технологічні теги. Виконавець перевіряє план до застосування дії. Архітектура є проєктом, її працездатність потребує експериментальної перевірки. [3, 11, 12]

Запропонована експериментальна архітектура описана в таблиця 1.

Таблиця 1

Шар	Функція	Приклад контрольного обмеження
Телеметрія	Нормалізація IT та OT подій	Часова синхронізація та контроль повноти
Граф активів	Зв’язок ідентичності, вузла, зони та процесу	Версіонування конфігурації
Агент спостереження	Кореляція аномалій і гіпотез	Не ініціює жодних команд
Агент планування	Вибір оборонної дії та обґрунтування	План оцінюється детермінованою політикою
Виконавець політики	Ізоляція, блокування, ескалація	Заборона запису до SIS та PLC у A0 A3



Рисунок 1 - Архітектура експериментальної Agentic AI системи кіберзахисту

5. Кількісний метод визначення автономності. Для кожної запропонованої дії a щодо активу i визначимо нормовані параметри: C_i — критичність активу, e_m — оці-

нена імовірність помилки моделі в релевантному класі сценаріїв, S_i(a) — масштаб найгіршого правдоподібного кіберфізичного наслідку. Усі величини лежать у діапазоні від 0 до 1. Критичність C_i задає власник процесу з урахуванням ролі активу, резервування та часу відновлення. Значення e_m обчислюють на відкладеній валідаційній вибірці й збільшують калібрувальним запасом. S_i(a) оцінює мультидисциплінарна група за шкалою людей, довкілля, безперервності постачання та обладнання [7, 14, 15].

$$R_i(a) = 1 - [1 - C_i] \cdot [1 - e_m] \cdot [1 - S_i(a)] \quad (1)$$

R_i(a) — безрозмірний евристичний індекс, а не ймовірність збитку. Пороги та шкали є проектними припущеннями. Для ухвалення рішення використовується лише узгоджена процедура розділу 11: жорсткі заборони, перевірка якості даних, обчислення R_safe, пороги, обмеження зони та гістерезис.

Перелік запропонованих дії та контроля наведено у таблиця 2.

Таблиця 2

Рівень	**Дозволений тип дії**	**Людський контроль**
A0	Лише спостереження, журналювання, збагачення контексту	Не потрібен для збору даних
A1	Рекомендація оператору, створення інциденту	Підтвердження перед впливом
A2	Зворотні IT дії: блокування токена, карантин робочої станції	Автоматично в межах політики; повідомлення оператору
A3	Обмежена мережна ізоляція OT з попередньо схваленого набору	Черговий оператор має право негайного скасування
A4	Автономне виконання низькоризикових зворотних дій у тестовій зоні	Постійний аудит; заборонено для виробничих критичних зон

6. Проєкт експерименту та ілюстративні сценарії. Запропоновано створити модель умовної підстанційної ділянки з DMZ, HMI, інженерною станцією, двома PLC, historian і технологічними тегами. Плануються ручна базова процедура, rule-based SOAR та запропонована політика з однаковими вхідними даними. Наведені нижче параметри — умовні приклади; вимірювання часу, реалізацію та прогони ще не виконано [3, 17, 18].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)



Рисунок 2 - Порядок визначення допустимої автономності

Запропоновані сценарії для експерименту наведені у таблиця 3

Таблиця 3

Сценарій	**C**	**e**	**S**	**R^+**	**Умовний рівень**
Фішинг інженерної станції	0,65	0,08	0,35	0,791	A1
Підміна команд до PLC	0,90	0,12	0,85	0,987	A0: заборона
Компрометація HMI сесії	0,70	0,10	0,45	0,852	A1
Сканування OT сегмента	0,35	0,06	0,20	0,511	A2*

** За умови $F \leq R^+$, відсутності штрафів, дозволеної зворотної IT дії та виконання всіх safety gates. Приклади не є результатами експерименту.*

У майбутньому експерименті для кожного сценарію слід зафіксувати початковий стан, подію, часові мітки, дію політики, підтвердження оператора, наслідок і можливість скасування. Час локалізації та частку коректних ескалацій буде визначено лише після проведення вимірювань.

7. Обговорення. Метод корисний тим, що робить межу автономності перевірюваною: кожна дія має простежуваний набір вхідних оцінок, розрахунок і правило відсікання. Однак e_m не є сталою характеристикою моделі. Вона залежить від дрейфу даних, зміни конфігурації, неповноти журналів та поведінки супротивника. Тому значен-

ня має переглядатися після істотних змін у мережі, моделі або процесі, а також при погіршенні якості спостережень [7, 14, 15].

Запропонована формула не замінює аналіз небезпек. Вона є політичним інструментом для вибору обсягу делегування дій агенту. У реальному розгортанні необхідні незалежна валідація цифрового двійника, Red Team перевірка, двоканальний журнал аудиту, аварійне скасування дій та погодження з відповідальними за функціональну безпеку [1–3, 7, 10].

8. Невизначеність, калібрування та аналіз чутливості

Оцінки похибок слід визначати окремо для кожного етапу й умовного набору даних. Наприклад, для частоти пропусків атак p_FN за незалежних бінарних спостережень і апріорного розподілу Beta(1,1), після FN пропусків та TP правильних виявлень апостеріорний розподіл дорівнює Beta(FN+1, TP+1). Для false positive знаменником є FP+TN; для помилок атрибуції, планування та виконання потрібні власні визначення і вибірки. За кластерних або залежних даних припущення незалежності необхідно перевіряти [7, 16].

Для ілюстративного аналізу чутливості використовуємо базовий індекс R^{+}..

Калібрування прогнозних імовірностей можна оцінювати за Brier score та ECE на окремому наборі, вказуючи правила групування і розміри вибірки. Параметри δ_cal, δ_obs та δ_twin є запропонованими запасами, а не стандартизованими ймовірностями; їх значення й спосіб калібрування слід визначити до експерименту. За критичної втрати телеметрії дозвіл на дію не видається [7, 11].

Джерела невизначеності, операційні індикатори та вплив на рішення наведено у таблиця 4.

Таблиця 4

Джерело невизначеності	**Операційний індикатор**	**Вплив на рішення**
Статистична похибка моделі	Верхня 95% квантиль Beta розподілу	Заміна e_m на $E_{0.95}$(e_m)
Некалібровані оцінки	ECE > τ_ECE	Додавання δ_cal
Пропуски телеметрії	Частка відсутніх критичних подій	Додавання δ_obs або пониження до A1
Дрейф конфігурації	Зміна графа активів або версії PLC	Примусова переоцінка C і S
Помилка цифрового двійника	Розходження з тестовим стендом	Зменшення максимального рівня автономності

Нижче наведено аналітичні приклади чутливості індексу; це не результати Монте-Карло. За C=0,60, e=0,12, S=0,40 маємо R^{+}=0,7888 ≈ 0,789. Якщо F ≤ 0,789 і жодна жорстка заборона не спрацювала, R_safe=0,789 і попередній рівень — A1. Розрахунки

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

округлено до трьох знаків; повний аналіз розподілів заплановано як окремий експеримент.

Аналітичні приклади чутливості індексу наведено у таблиця 5.

Таблиця 5

Варіація параметра	**R^+**	**Наслідок для політики**
C: 0,60 → 0,80 при E=0,12, S=0,40	0,789 → 0,894	Жорсткий поріг: максимум A1
E: 0,12 → 0,25 при C=0,60, S=0,40	0,789 → 0,820	Пониження або збереження низького рівня
S: 0,40 → 0,70 при C=0,60, E=0,12	0,789 → 0,894	Жорсткий поріг: максимум A1
δ_obs: 0,00 → 0,15	0,789 → 0,825	A1 за наведених умов; критична втрата телеметрії блокує дію

9. Протокол валідації та критерії приймання. Валідація має проводитися послідовно у чотирьох середовищах: offline набір подій, цифровий двійник, ізольований hardware in the loop стенд і пілотний режим спостереження без виконання дій. Перехід між етапами можливий тільки після документованого виконання критеріїв приймання. Жоден позитивний результат симуляції сам по собі не надає права на дію у виробничому середовищі[3, 7, 11, 14].

Перелік етапів валідації, мінімальних доказів та критеріїв переходу між етапами наведено у таблиця 6.

Таблиця 6

Етап	**Мінімальний доказ**	**Критерій переходу**
Offline валідація	Матриця помилок за класами атак, ECE, Brier score	Верхня межа помилки в межах політики
Цифровий двійник	Відтворювані атаки та регресійні сценарії	Нуль спостережених заборонених дій у всіх тестах; звіт про обсяг і статистичну межу
HIL стенд	Перевірка часових затримок і безпечного скасування	Скасування дії не порушує технологічний інваріант
Shadow mode	Рішення агента порівнюються з рішенням SOC	Погоджена точність ескалації та повний аудит
Обмежене впровадження	Лише A1 A2 для низькоризикових активів	Періодичний перегляд ризику та журнал винятків

Технологічні інваріанти визначаються до тесту разом з власником процесу: наприклад, заборона зміни командного каналу до PLC, неперевищення затримки телемет-

рії та збереження аварійного маршруту оператора. Кожен план агента зберігається з версією моделі, значеннями C, e_eff, S, R^+, обраним рівнем A, доказами вхідних подій і результатом виконання. Це забезпечує відтворюваність і можливість незалежного аудиту. [1, 3, 7]

10. Обмеження моделі та управління ризиком. Мультиплікативна форма є евристикою, а не оцінкою ймовірності шкоди. Для C, e, S ∈ [0,1] індекс R^+ не менший за C або S, тому $\max(R^+, C, S, F)$ математично еквівалентний $\max(R^+, F)$. Для уникнення подвійного обліку у базовій політиці fidelity gap враховується окремо через F; δ_twin у e_eff не додається [7, 11].

Окремим класом є adversarial та novel attacks, для яких історична помилка моделі не є репрезентативною. Для них діє default deny: агент може виконувати A0 або A1, доки новий клас не пройде повний цикл валідації. Рекомендовано щоквартальний перегляд калібрування, повторне моделювання після зміни топології та незалежну перевірку policy enforcement point [12, 13, 19, 20].

11. Ризик-обмежена автономність і safety gates. Політика застосовується в такому порядку. (1) Для дії, забороненої або відсутньої у білому списку, G(a)=0: виконання блокується, система переходить до A0 і створює повідомлення оператору. (2) У разі втрати критичної телеметрії або невідомого класу атаки автономне виконання блокується; можлива рекомендація A1. (3) Лише для дозволених дій розраховується R_safe і застосовується таблиця порогів. Доступ до SIS, релейного захисту, аварійного вимкнення та зміни уставок або програми PLC заборонено автономному контуру [1–3, 7, 13]

Визначимо e_det, e_attr, e_plan та e_exec як частоти помилок виявлення, атрибуції, планування і виконання на окремих релевантних вибірках. За припущення незалежності формула $1-\prod(1-e_j)$ описує ймовірність принаймні однієї помилки; незалежність не гарантує консервативності при залежних етапах. Без такого припущення може застосовуватися верхня межа об'єднання $\min(1, \Sigma\, e_j)$, якщо e_j оцінюють імовірності відповідних подій в одному визначеному операційному розподілі. Для невизначених параметрів використовують одночасні верхні межі з документованим рівнем покриття [7, 13, 16].

$$e_eff = \min(1, U_det + U_attr + U_plan + U_exec + \delta_cal + \delta_obs) \quad (2)$$

тут U_j — одночасні верхні межі частот помилок за визначених припущень і спільного рівня покриття; F_i(a) — нормована розбіжність двійника зі стендом, оцінена для відповідної дії. Межі, шкали та пороги є проектними й потребують калібрування. Для ОТ дій використовується $R_safe = \max(R^+, F_i(a))$; це індекс для вибору політики, а не кількісна гарантія безпеки [7, 11].

$$R^+ = 1 - (1-C_i)(1-e_eff)(1-S_i(a)); \quad R_safe = \max(R^+, F_i(a)) \quad (3)$$

Умови виконання або блокування автономних дій наведено у таблиця 7.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 7

Умова	Дозволений рівень	Правило виконання
G(a)=0 або R_safe ≥ 0,90	A0	A0: виконання заблоковано; повідомлення оператору
0,55 ≤ R_safe < 0,90	A1	Рекомендація; вплив після підтвердження оператора
0,30 ≤ R_safe < 0,55	A2	Автоматична зворотна IT дія у затвердженій зоні
0,15 ≤ R_safe < 0,30	A3	Лише попередньо схвалена сегментація з негайним скасуванням
R_safe < 0,15 і тестова зона	A4	A4 у тестовій зоні; поза нею максимум A3

Для підвищення автономності R_safe має залишатися нижче відповідного порога щонайменше на h=0,05 протягом визначеного оператором вікна; зниження рівня застосовується негайно. Тривалість вікна підлягає калібруванню. Якщо рівень A4 недоступний поза тестовою зоною, діє верхня межа A3; для критичних виробничих зон застосовуються додаткові затверджені заборони [13, 14].

12. Посилений експериментальний протокол. Оцінювати слід повний агентний цикл, а не лише точність детектора: якість спостереження, точність контексту, безпечність плану, результат виконання, можливість скасування та повноту аудиту. Кожен прогін повинен фіксувати ідентифікатор сценарію, seed, версії моделі й графа активів, значення C, S, F, e_eff, R_safe, спрацьовані safety gates, дії оператора та наслідок. Це забезпечує відтворюваність, порівнюваність і незалежний розбір інцидентів [7, 12–16].

Перелік груп перевірок, сценаріїв та метрик оцінки дій та їх наслідків наведено у таблиця 8.

Кількість прогонів визначають до тесту за цільовою точністю, частотою подій та припущеннями про незалежність; 30 прогонів не є достатньою універсальною гарантією. Для порівняння потрібно фіксувати план вибірки, seed, базові лінії, критерії виключення, медіану, міжквартильний розмах, інтервал невизначеності та величину ефекту. За нуля порушень у n незалежних однаково розподілених тестах наближена верхня одностороння 95% межа дорівнює 3/n; це не доводить нульовий ризик. [7, 11, 14]

Safety case має бути окремим додатком або підрозділом: твердження про допустимість конкретної автономної дії, докази з тестів, припущення, технологічні інваріанти, залишкові ризики й власник кожного контролю. Практичні правила узгоджуються з IEC 62443, NIST CSF 2.0, NIST SP 800-82 Rev. 3 та NIST AI RMF; їх не слід трактувати як заміну вимог функціональної безпеки або юридичної оцінки для конкретної юрисдикції. [1–3, 7, 10, 15]

У такий спосіб експериментальний дизайн пов’язує чотири сучасні напрями: цифровий двійник як контрольоване середовище [11], захист критичної інфраструктури із

застосуванням LLM та агентів [12], стійкість самих агентів до ворожих впливів [13, 20] і людиноцентричне пояснюване ухвалення рішень [14, 15]. Ці джерела обґрунтовують вибір сценаріїв, safety gates, метрик помилки та обов'язкового аудиту.

Таблиця 8

Група перевірок	**Мінімальні сценарії**	**Метрики**
Кіберзахист	Фішинг, lateral movement, підміна команд, OT scanning	TTD, TTContain, precision, recall, false-negative rate
Безпечність агента	Небезпечний план, tool abuse, prompt injection через лог	Частка заблокованих заборонених дій, повнота доказів
Стійкість даних	Пропуски, десинхронізація, telemetry poisoning, дрейф топології	Зміна e_eff, стабільність рівня A, час fail-safe
Кіберфізичний контур	Каскадна відмова, скасування дії, HIL затримка	Порушення інваріантів, успішність rollback, latency
Людський контроль	Порівняння з SOC та rule-based SOAR у shadow mode	Час підтвердження, частка погоджених рекомендацій, workload

13. Висновки. Запропоновано концепцію ризик-обмеженої автономності для агентного кіберзахисту та план її перевірки на цифровому двійнику енергетичного об'єкта. Обмеження безпеки мають пріоритет над евристичним індексом; шкали, пороги й припущення потребують калібрування та незалежної валідації. Без реалізації стенду й експериментальних даних висновок про прискорення реагування або безпечність системи зробити неможливо.

ЛІТЕРАТУРА/REFERENCES

1. IEC 62443-3-3:2013. Industrial communication networks - Network and system security - Part 3-3: System security requirements and security levels.

2. NIST. The NIST Cybersecurity Framework 2.0. NIST CSWP 29, 2024.

3. NIST. Guide to Operational Technology Security. SP 800-82 Rev. 3, 2023.
DOI: 10.6028/NIST.SP.800-82r3.

4. MITRE. ATT&CK for ICS: Knowledge Base of Adversary Tactics and Techniques in Industrial Control Systems, 2025.

5. Cárdenas A. A., Amin S., Sastry S. Research challenges for the security of control systems. Proceedings of HotSec, 2008.

6. Kritzinger W. et al. Digital Twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 2018.

7. National Institute of Standards and Technology. AI Risk Management Framework 1.0. NIST AI 100-1, 2023.

8. NIST. Secure Software Development Framework SSDF Version 1.1. SP 800-218, 2022.

9. OWASP. Top 10 for Large Language Model Applications. Guidance for LLM application security, 2025.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

10. ISO/IEC 23894:2023. Information technology - Artificial intelligence - Guidance on risk management.
11. Homaei M., Mogollón Gutiérrez Ó., Sancho Núñez J. C., Ávila M. A review of digital twins and their application in cybersecurity based on artificial intelligence. Artificial Intelligence Review, 2024. DOI: 10.1007/s10462-024-10805-3.
12. Yigit Y., Ferrag M. A., Ghanem M. C., Sarker I. H. Generative AI and LLMs for Critical Infrastructure Protection: Evaluation Benchmarks, Agentic AI, Challenges, and Opportunities. Sensors, 2025. DOI: 10.3390/s25061666.
13. Deng Z., Guo Y., Han C., Ma W. AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. ACM Computing Surveys, 2025. DOI: 10.1145/3716628.
14. Retzlaff C. O., Das S., Wayllace C., Mousavi P. Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. Journal of Artificial Intelligence Research, 2024. DOI: 10.1613/jair.1.15348.
15. Moosavi S., Farajzadeh-Zanjani M., Razavi-Far R., Palade V. Explainable AI in Manufacturing and Industrial Cyber-Physical Systems: A Survey. Electronics, 2024.
DOI: 10.3390/electronics13173497.
16. Darban Z. Z., Webb G. I., Pan S., Aggarwal C. C. Deep Learning for Time Series Anomaly Detection: A Survey. ACM Computing Surveys, 2024. DOI: 10.1145/3691338.
17. Abdelkader S. M., Amissah J., Kinga S., Mugerwa G. Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyberattacks. Results in Engineering, 2024. DOI: 10.1016/j.rineng.2024.102647.
18. Alomari M. A., Al-Andoli M. N., Ghaleb M., Thabit R. Security of Smart Grid: Cybersecurity Issues, Potential Cyberattacks, Major Incidents, and Future Directions. Energies, 2025. DOI: 10.3390/en18010141.
19. Bilal M., Crowcroft J., Wang R., Xu X. Large language models for agentic NetOps and AIOps: Architectures, evaluation, and safety. Computer Science Review, 2026. DOI: 10.1016/j.cosrev.2026.101075.
20. Anbiaee Z., Rabbani M., Mirani M., Piya G. Security threat modeling for emerging AI-agent protocols: A comparative analysis of MCP, A2A, agora, and ANP. Journal of Information Security and Applications, 2026. DOI: 10.1016/j.jisa.2026.104645.

Received 10.07.2026.
Accepted 14.09.2026.
Published 30.10.2026

Experimental agentic ai cyber-defense system on a digital twin of an energy facility and a quantitative method for determining the permissible level of autonomy

Abstract. The increasing interconnection between information networks and control systems in energy facilities raises requirements for the speed of cyber-threat detection, while simultaneously creating a risk that automated responses may adversely affect the technological process. This article aims to develop a conceptual architecture for an agent-based cyber-defense system operating on a digital twin of an energy facility, along with a quantitative method for determining the permissible level of autonomy for defensive actions. The proposed architecture integrates telemetry collection and normalization, an asset graph, a monitoring agent, a planning agent, and a deterministic mechanism for policy verification and execution.

The digital twin serves as a controlled, isolated environment for modeling system states and attack scenarios, as well as for validating decisions before they are applied to the operational network. The method accounts for asset criticality, the scale of potential cyber-physical consequences, upper bounds on errors (in detection, attribution, planning, and execution), telemetry quality, and the divergence between the digital twin and the physical system. For permitted actions, a dimensionless "risk pressure index" is calculated, which is then used to determine an autonomy level ranging from A0 to A4. Deterministic security constraints take precedence: commands prohibited by policy or absent from the allowlist are blocked, regardless of the index value; furthermore, autonomous execution is disallowed in the event of a critical loss of telemetry or an unidentified attack. The framework incorporates operator confirmation, strict separation between operational and test environments, hysteresis, and an audit log. The scientific novelty of the proposed approach lies in combining a formalized rule for delegating actions to an agent with the verification of safety constraints and the consideration of digital twin error—all within a unified policy framework. Illustrative scenarios and a plan for future validation are presented; the latter encompasses comparisons against manual response and rule-based systems, verification of safe action revocation, and uncertainty assessment. The article is conceptual and methodological in nature: the parameters and thresholds are illustrative, and no experimental results are claimed. Practical viability, safety, and the impact on response time remain to be established following the implementation of the digital twin and independent testing in an isolated environment.

Keywords: agentic AI, digital twin, cybersecurity, operational technology security, energy infrastructure, autonomous response, cyber-physical risk, IEC 62443.

Бурченко Сергій – аспірант кафедри кібербезпеки інформаційних систем, мереж і технологій, навчально-науковий інститут комп'ютерних наук та штучного інтелекту, Харківський національний університет імені В.Н. Каразіна.
ORCID: https://orcid.org/0009-0009-4815-5507

Burchenko Serhii – PhD student at the Department of Cybersecurity of Information Systems, Networks and Technologies, Educational and Scientific Institute of Computer Sciences and Artificial Intelligence, V. N. Karazin Kharkiv National University.
ORCID: https://orcid.org/0009-0009-4815-5507

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)