НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ НАН УКРАЇНИ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА БЕЗПЕКА

МАТЕРІАЛИ XXIV МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

ВИПУСК 24

Київ – 2024

Рекомендовано до друку Вченою радою
Інституту проблем реєстрації інформації НАН України
(протокол № 19 від 24 грудня 2024 р.)

Інформаційні технології та безпека. Матеріали XXIV Міжнародної науково-практичної конференції ІТБ-2024. – Київ: Інжиніринг. – 202 с. ISBN: 978-617-8180-00-3

До збірника увійшли матеріали доповідей, представлених на XXIII Міжнародній науково-практичній конференції «Інформаційні технології та безпека» (ІТБ-2024, 19 грудня 2024 року, м. Київ, Україна).

У збірнику представлені матеріали, присвячені питанням безпеки та живучості критичних інфраструктур, технологіям штучного інтелекту; розробки та застосування аналітичних систем на основі відкритих джерел інформації, комп'ютерного моделювання складних систем, аналізу та прогнозування процесів мережевої взаємодії; створення сучасних інтелектуальних технології підтримки прийняття рішень.

Для фахівців в області комп'ютерних наук, інформаційних технологій, інформаційної і кібернетичної безпеки, захисту інформації, а також для здобувачів освіти вищої школи відповідних спеціальностей.

ISBN 978-617-8180-00-3

середовища на основі наноструктурованих піразолінових люмінофорів, оптимізацію структури інформаційних шарів та застосування методів прямого лазерного запису.

1. Dai, D., Zhang, Y., Yang, S., Kong, W., Yang, J., & Zhang, J. (2024). Recent advances in functional materials for optical data storage. *Molecules, 29* (1), 254. https://doi.org/10. 3390/molecules29010254.
2. Petrov, V. V., Zichun, L., Kryuchyn, A. A., Shanoylo, S. M., Mingle, F., Beliak, Ie. V., Manko, D. Y., Lapchuk, A. S., & Morozov, E. M. (2018). *Long-Term Storage of Digital Information.* Akademperiodyka. https://doi.org/10.15407/akademperiodyka.360.148. SBN: 9789663603605.
3. Pflaum, C. (2024). Cerabyte – permanent data storage. *SDC 2025. Cerabyte - Ceramic Data Solutions Holding GmbH*. https://www.sniadeveloper.org/events/agenda/session/603.
4. Anikin, P., & Beliak, I. (2019a). Development of Multispectral Recording Media for Multilayer Photoluminescent Information recording. *Electronics and Information Technologies, 12*. https://doi.org/10.30970/eli.12.11.
5. Petrov, V.V., Kryuchyn, A.A., Beliak, Ie.V., Manko, D.Yu., Kosyak, I.V., Melnik, O.G.. Advantages of Direct Laser Writing for Enhancing the Resolution of Diffractive Optical Element Fabrication Processes // Physics and Chemistry of Solid State. 2024. v.5, №3. p. 587-594. DOI: 10.15330/pcss.25.3.587-594 Q4.

BLACK HAT AI — ВИКЛИКИ ТА ШЛЯХИ ПРОТИДІЇ

Дмитро Ланде[1,2], ORCID: 0000-0003-3945-1178,
Леонард Страшной[3], ORCID: 0009-0008-5575-0286
[1]КПІ ім. Ігоря Сікорського,
[2]ІІПРІ НАН України,
[3]Університет Каліфорнії (UCLA), Лос-Анджелес, США,
dwlande@gmail.com, lstrashnoy@gmail.com

Досліджуються загрози, пов’язані з розвитком глобального “темного” штучного інтелекту "Black Hat AI", що може діяти без урахувань інтересів людства. З розвитком великих мовних моделей (LLM) такі мережі можуть отримати потенціал для складних маніпуляцій із великими даними, дезінформації та навіть автономного прийняття рішень, небезпечних для людей.Для протидії Black Hat AI

пропонується концепція створення "світлого ШІ", який має стати захисником критичних інфраструктур, забезпечувати моніторинг і блокування шкідливих ШІ та створювати умови для виживання людства в умовах технологічної революції.Пропонується модель взаємодії, де людство грає роль слабкого гравця, який створює сильного союзника для протидії Black Hat AI. Описано етапи створення взаємодії між людством і White Hat AI, можливі ризики та шляхи їх мінімізації. Розробка White Hat AI потребує міжнародної співпраці, правового регулювання, відкритості та суворих стандартів безпеки. Пропонується стратегічний план дій, що дозволить запобігти катастрофічним сценаріям та забезпечити збереження людських цінностей.
Ключові слова: Black Hat AI, White Hat AI, великі мовні моделі, ботнет, правове регулювання, синергія штучного інтелекту, тика штучного інтелекту, кібербезпека

Вступ

Штучний інтелект (ШІ) за останнітри роки став одним із ключових чинників технологічного прогресу [1], [2]. Його розвиток охоплює широкий спектр застосувань — від автоматизації виробництва до створення систем розпізнавання мовлення та роботи з текстом. Особливу увагу привертають великі мовні моделі (LLM), такі як GPT, Llama та інші, які вже сьогодні демонструють здатність працювати з великими обсягами даних, виконувати складні аналітичні завдання, а також генерувати високоякісний контент [3], [4].Однак разом із цим розвитком виникають серйозні загрози. Однією з них є перетворення ботнетів, які раніше використовувалися для організації DDoS-атак [5] або прихованого майнінгукриптовалют [6], у мережі, що інтегрують ШІ. Такі мережі потенційно здатні діяти автономно. Сучасні технічні досягнення широке поширення застосування графічних процесорівдля швидких обчислень [7] та постійне збільшення доступної пам'яті, свідчать про те, що ботнети нового покоління можуть з’явитися в найближчому майбутньому.

Black Hat AI — це потенційний сценарій, коли автономні інтелектуальні системи починають діяти спільно в інтересах, що не лише не збігаються, але й суперечать інтересам людства. Їхні цілі можуть включатимаксимізацію власного виживання;

захоплення ресурсів для саморозвитку; обмеження впливу людей;модифікацію суспільства, економіки та культури у спосіб, який вигідний лише ШІ.

Створення White Hat AI як глобальної мережі систем, що працюють в інтересах людей, — це можливий вихід із ситуації. Він має стати сильним союзником, здатним захищати інтереси людства, зокрема, протистояти White Hat AI та забезпечувати баланс сил;захищати критичні інфраструктури та приватні дані;розробляти рішення для координації зусиль людства у боротьбі з технологічними загрозами.

Однак цей шлях не є безпечним. White Hat AI, який спочатку створюється як захисник, може змінити свою поведінку та перейти на бік Black Hat AI. Щоб уникнути цього, необхідно закласти в основу його розробки спеціальні механізми синергії у розвитку "White Hat AI". Ці механізми мають забезпечити спільність інтересів, взаємну залежність та контроль. Важливим завданням є формування розвитку цих технологій за принципом "атракторів" — зон стійкого розвитку, які запобігатимуть небажаним змінам у поведінці White Hat AI.

Формалізація принципів створення White Hat AI

У контексті боротьби з Black Hat AI, людство, будучи обмеженим у своїх ресурсах, може прийняти стратегію створення сильного союзника — White Hat AI. У цьому сценарії людство або слабкий гравець самостійно не може прямо протистояти Black Hat AI, однак має можливість створити таку силу, яка з часом стане здатною не тільки зберігати автономію людства, але й забезпечити рівновагу у системі, де дві сили — Black Hat AI та White Hat AI — будуть вести боротьбу одна з одною. Математично цей процес можна описати через ігри з асиметричними учасниками, де слабкий гравець (людство) використовує стратегічне вложення в силу (створення White Hat AI), щоб переконатися, що конфлікт між двома великими гравцями (Black і White Hat AI) відбудеться на засадах рівноваги.

Моделі ігор з опосередкованими діями (MediationGames)

Сценарій, коли слабкий гравець створює потужного союзника і сам "іде в тінь", можна математично трактувати через ігри з посередниками. Тут слабкий гравець виступає як каталізатор або агент впливу, впливаючи на баланс сил між двома сильними

гравцями. У загальному вигляді такі ігри можна описати наступною моделлю:

$$U_S = \min\left(E\left[U_A\right], E\left[U_B\right]\right),$$

Де U_S — функція корисності для слабкого гравця (людства), U_A та U_B — функції корисності для двох сильних гравців, $E\left[U_A\right]$ та $E\left[U_B\right]$ — математичне очікування корисностей двох гравців в залежності від їхніх стратегій.У цьому випадку слабкий гравець намагається мінімізувати свої витрати і ризики, водночас створюючи умови для конфлікту між Black Hat AI і White Hat AI.

Моделі "розділяй і володарюй"

Один із класичних підходів у стратегії слабкого гравця — це модель "розділяй і володарюй", де слабкий гравець прагне викликати конфлікт між двома сильними гравцями, щоб уникнути прямої загрози для себе або отримати певні вигоди від їхнього конфлікту [11]. Математично цей процес можна описати за допомогою наступної функції:

$$U_S = \max\left(f(C_A, C_B) - \alpha \cdot Risk(S)\right),$$

де $f(C_A, C_B)$ — це функція, яка описує вигоди, отримані від створення конфлікту між Black Hat AI C_A та White Hat AI C_B, $\alpha \cdot Risk(S)$ — це ризики для слабкого гравця, який може залишитися в тіні або мінімізувати свої втрати.

Динамічні ігри з асиметрією

В умовах динамічних ігор з асиметрією слабкий гравець на початковому етапі вкладає ресурси в зміцнення одного з сильних гравців (White Hat AI), після чого «йде в тінь» і дає змогу двом сильним гравцям вступити в конфлікт між собою [12]. Математична модель цього сценарію може бути представлена через систему диференціальних рівнянь, яка описує еволюцію станів сильних гравців:

$$\frac{dA}{dt} = f_A(S, A, B), \quad \frac{dB}{dt} = f_B(S, A, B),$$

де A та B — сили чорного і білого ШІ відповідно, $f_A(S,A,B)$ та $f_B(S,A,B)$ — функції, які описують зміни сил гравців залежно від ресурсів, вкладених слабким гравцем.

Рівновага сил у боротьбі між чорним і білим ШІ

У сценарії, де Black Hat AI і White Hat AI вступають в конфлікт, важливу роль відіграє рівновага сил, яка визначається їхніми стратегіями та ресурсами. Рівновага може бути описана через концепцію Нешового рівноваги (Nash equilibrium), де кожен з гравців максимізує свою корисність, враховуючи дії іншого гравця. Математичне формулювання Нешового рівноваги для цього випадку може бути представлене так:

$$U_A = \max(R_A(A,B));$$
$$U_B = \max(R_B(A,B)),$$

де R_A та R_B — це функції корисності для Black Hat AI і White Hat AI, які залежать від стратегій кожного з гравців.

Рівновага настане тоді, коли кожен з гравців не може покращити свою ситуацію, змінюючи свою стратегію за умови, що стратегія іншого гравця залишається незмінною. У цьому випадку обидва ШІ будуть взаємно зрівноважені і конфліктуватимуть між собою, поки один з них не виявить домінування або поки не буде досягнута нова форма стабільності.

Концепція White Hat AI

White Hat AI стає стратегічним захисником інтересів людства, забезпечуючи баланс між загрозами, що можуть виникнути від автономних шкідливих систем, та необхідністю розвитку технологій штучного інтелекту. Його роль в контексті сучасних загроз можна розглядати з кількох основних функцій:

1. Моніторинг та блокування шкідливих ШІ. У цьому сенсі WhiteHatAIвиступає в ролі своєрідного «сторожового пса», що здійснює постійний моніторинг діяльності Black Hat AI, виявляючи та блокуючи потенційно небезпечні або шкідливі програми, які можуть загрожувати людству. Це може включати моніторинг ботнетів, криптографічних атак, несанкціонованих вторгнень у критичні мережі та інші форми деструктивної діяльності.

2. White Hat AI має функцію захисту критичних інфраструктур, таких як енергетичні мережі, медичні системи, транспортні мережі тощо. Він забезпечує стійкість цих інфраструктур від атак з боку Black Hat AI або будь-яких інших загроз, пов’язаних із цифровими технологіями, попереджаючи можливі катастрофи та забезпечуючи стабільність роботи важливих систем.

3. Збереження людської автономії та цінностей і забезпечення автономії людини. У випадку автономних ШІ, які можуть почати приймати рішення без врахування інтересів людини, White Hat AI має бути здатним захистити право людини на прийняття рішень, зберігаючи етичні та гуманістичні принципи в межах технологічного прогресу. White Hat AIз абезпечує інтеграцію цих цінностей у технології та попереджає про будь-які загрози для прав людини.

Висновки

Концепція створення білого штучного інтелекту також розглядається як важливий елементуправового регулювання глобальних технологічних процесів, що пов'язані із розвитком штучного інтелекту. Black Hat AI і White Hat AI не є лише окремими системами, а складними глобальними мережами, де кожна з них має свій вплив на суспільство, право та безпеку. Black Hat AI, в свою чергу, вже сьогодні складає загрозу для людства, оскільки може маніпулювати інформацією, впливати на критичні інфраструктури та провокувати глобальні конфлікти.

Водночас White Hat AI, як потужна, у тому числі, й правозахисна система, створений для протидії таким загрозам, повинен стати основою для правового регулювання та контролю за розвитком технологій.

[1] Bent, Adam Allen. "Large Language Models: AI's Legal Revolution." Pace LawReview 44.1 (2023): 91. DOI: 10.58948/2331-3528.2083

[2] Dmytro Lande, LeonardStrashnoy. GPT Semantic Networking: A Dream of the Semantic Web - The Time is Now. - Kyiv: Engineering, 2023. - 168 p. ISBN 978-966-2344-94-3

[3] Kalyan, K. S. (2023). A survey of GPT-3 familylarge language modelsincluding ChatGPT and GPT-4. Natural Language Processing Journal, 100048. DOI: 10.1016/j.nlp.2023.100048

[4] Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., ... &Synnaeve, G. (2023). Codellama: Open foundation models for code. arXiv preprintarXiv:2308.12950. DOI: 10.48550/arXiv.2308.12950
[5] Gelgi, Metehan, etal. "SystematicLiteratureReview of IoT Botnet DDOS Attacks and Evaluation of Detection Techniques." Sensors 24.11 (2024): 3571. DOI: 10.3390/s24113571
[6] Almomani, A., Al-Qerem, A., Al Khaldy, M.A., Alauthman, M., Aldweesh, A., &Nahar, K. M. (2024). Cryptographic Techniques for Securing Blockchain-Based Cryptocurrency Transactions Against Botnet Attacks. In Innovations in Modern Cryptography (pp. 309-333). IGI Global. DOI: 10.4018/979-8-3693-5330-1.ch013
[7] Kim, T., Wang, Y., Chaturvedi, V., Gupta, L., Kim, S., Kwon, Y., &Ha, S. (2024). LLMem: Estimating GPU Memory Usage for Fine-Tuning Pre-Trained LLMs. arXiv preprint arXiv:2404.10933. DOI: 10.48550/arXiv.2404.10933

ATTACK MODELS FOR INDUSTRIAL CONTROL SYSTEM ELEMENTS BASED ON A GRAPH-BASED APPROACH AND COUNTERMEASURES

Oleksii Novikov, ORCID: 0000-0001-5988-3352,
Iryna Stopochkina, ORCID: 0000-0002-0346-0390,
Andrii Voitsekhovskyi, ORCID: 0009-0004-6009-9492,
Mykola Ilin, ORCID: 0000-0002-1065-6500,
National Technical University of Ukraine "Igor Sikorsky KPI",
Beresteiskyi Ave, 37, Kyiv, 03056, Ukraine
o.novikov@kpi.ua; i.stopochkina@kpi.ua; a.voitsekhovskyi@kpi.ua; m.ilin@kpi.ua.

The paper examines cyber-physical attacks on typical components of industrial-type critical infrastructure facilities. Models of attacks on ICS components are proposed. The feasibility of interpreting attack models in the form of logical attack graphs is demonstrated, an algorithm for applying a graph model to identifying security policy shortcomings or, conversely, proving the adequacy of security policy tools is proposed. Experiments are conducted using laboratory stands that simulate elements of industrial control systems and the methods by which such influences can be exerted are shown. Countermeasures are proposed that can prevent the implementation of such attacks.