Для підвищення стійкості слід застосовувати принцип мінімально необхідного доступу: пристрій взаємодіє лише з ресурсами, потрібними для його функцій, а адміністративні інтерфейси ізолюються від публічних мереж і захищаються багатофакторною автентифікацією. Європейський підхід до критичних суб'єктів акцентує безперервність функцій, урахування взаємозалежностей і здатність до відновлення після інцидентів [4].

Отже, мережева безпека IoT-пристроїв є базовою умовою надійності кіберфізичних систем розумного міста. Запропонований підхід охоплює автентифікацію, шифрування, сегментацію, моніторинг, журналювання та реагування. Подальші дослідження доцільно спрямувати на автоматизоване виявлення аномалій трафіку, адаптацію Zero Trust і моделі оцінювання ризиків для різних IoT-мереж.

1. Humayed A., Lin J., Li F., Luo B. Cyber-physical systems security - A survey. IEEE Internet of Things Journal. 2017. Vol. 4(6). P. 1802-1831.
2. Mosenia A., Jha N.K. A comprehensive study of security of Internet-of-Things. IEEE Transactions on Emerging Topics in Computing. 2017. Vol. 5(4). P. 586-602.
3. Khan M.A., Salah K. IoT security: Review, blockchain solutions, and open challenges. Future Generation Computer Systems. 2018. Vol. 82. P. 395-411.
4. Dritsas E., Trigka M. A survey on cybersecurity in IoT. Future Internet. 2025. Vol. 17(1). Article 30.

AI bots as a factor reducing the cyber resilience of virtual communities on social networking services

UDK 004.056:004.738.5(045)

Vadym Kolesnyk

Kharkiv National University of Radio Electronics, vadym.kolesnyk@nure.ua

The cyber resilience of virtual communities on social networking services is determined by their ability to maintain functionality, uphold trust among participants, ensure communicative continuity, adapt to disruptive influences, and recover from them. One factor reducing cyber resilience is AI bots, which can mimic human behavior, automatically generate messages, perpetuate specific narratives, provoke conflicts, and complicate the detection of coordinated inauthentic activity [1–2].

The urgency of the issue is heightened by the fact that AI is already considered one of the defining factors of the modern cyberthreat landscape. In particular, ENISA notes that AI is used to enhance social engineering, generate content, and increase the effectiveness of destructive campaigns [3]. In the case of virtual communities, the danger lies not only in the spread of misinformation but also in the gradual degradation of the socio-technical environment.

The goal is to identify sets of indicators that can be used to assess the impact of AI bots on the cyber resilience of virtual communities on social networking services.

Existing research has primarily focused on bot detection, disinformation analysis, automated moderation, network resilience, and the assessment of behavioral

anomalies [3]. Such approaches are important because they enable the identification of suspicious accounts, the classification of destructive content, and the detection of specific signs of coordinated influence. At the same time, their limitation lies in their fragmentary nature: they do not always reveal whether a community maintains its cyber resilience after exposure to AI bots.

It is advisable to view AI bots as a factor in the systemic destabilization of a virtual community, affecting the four interrelated dimensions of its cyber resilience (Fig. 1).

Fig. 1. Dimensions of a virtual community's cyber resilience

The communicative dimension encompasses an increase in the proportion of repetitive or formulaic posts, a decline in the depth of discussions, a rise in the number of contentious debates, the spread of toxic reactions, and a decrease in substantive reciprocity among participants. The behavioral dimension includes abnormal posting frequency, synchronized actions by groups of accounts, unnatural activity, sudden spikes in comments or reactions, and uniformity in communication patterns. Trust-related indicators are associated with a decline in real user participation, a suspicious increase in the number of new accounts, the fragmentation of discussions, and a weakening of social support within the community. Moderation-related indicators characterize an increase in the number of complaints, longer response times from moderators, the reappearance of destructive content, and an increase in the number of repeat violations.

Unlike approaches that assess only signs of the presence of bots or bot-like activity, the proposed approach focuses on evaluating their impact on the functioning of the community. This allows us to view AI bots not merely as a technical or informational risk, but as a cybersecurity factor that can alter the structure of interactions, participant behavior, communication channels, and the effectiveness of moderation mechanisms. In this context, structural stability is just one component of broader cyberresilience: a community may maintain formal activity but lose the trust of its members, engage in destructive communication, and lose its ability to recover.

A limitation of the proposed approach is the difficulty of reliably distinguishing between AI bots, ordinary automated accounts, and overly active real users. An

additional challenge is the uneven access to data across different platforms, particularly to moderation logs, interaction histories, and account metadata.

Further research should focus on building a model of a virtual community in which the proportion of AI bots, the intensity of their activity, the level of toxicity, and the speed of moderation vary. This will allow us to identify empirical threshold conditions under which AI bots do not yet disrupt community functioning, as well as critical states under which significant destructive effects, fragmentation of interactions, and a decline in cyberresilience occur.

1. Molodetska K. Information Influence on the Virtual Community: Implementation Features and Method of Detection in Social Internet Services / K. Molodetska, S. Veretiuk, I. Rahimova, S. Milevskyi, V. Khvostenko // 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). Ankara, Turkiye, 2023. P. 1–6. DOI: 10.1109/ISMSIT58785.2023.10305001.
2. Doshi J., Novacic I., Fletcher C., Borges M., Zhong E., Marino M. C., Gan J., Mager S., Sprague D., Xia M. Sleeper Social Bots: a new generation of AI disinformation bots are already a political threat. arXiv. 2024. DOI: 10.48550/arXiv.2408.12603.
3. ENISA. ENISA Threat Landscape 2025. Luxembourg: Publications Office of the European Union, 2025.

Аналіз атак витоку системних інструкцій у великих мовних моделях

УДК 004.056.5

Віктор Кольченко[1]

Національний університет «Львівська політехніка»,
[1]viktor.v.kolchenko@lpnu.ua

Стрімкий розвиток великих мовних моделей (LLM) та їх інтеграція в інтелектуальних агентів створили нові вектори загроз, серед яких витік системних інструкцій (Prompt Leakage) стає однією з найбільш критичних проблем безпеки. Ця вразливість полягає в ненавмисному розкритті прихованих налаштувань, які визначають логіку поведінки, обмеження безпеки та операційні параметри моделі [1]. В сучасній екосистемі ШІ системні інструкції перетворюються на найцінніший актив інтелектуальної власності, що створює прямі стимули для їх викрадення.

Мета дослідження полягає у проведенні аналізу атак на витік системних інструкцій, дослідженні внутрішніх механізмів їх реалізації для забезпечення безпеки інтелектуальної власності в екосистемах ШІ-агентів.

Системні інструкції у LLM функціонують як прихований шар управління, що задає роль моделі, її тон та межі дозволеної взаємодії. У сучасних агентних системах вони еволюціонують від простих текстових промптів до модульних пакетів навичок, які поєднують робочі процеси, використання інструментів та специфічні доменні знання. Сутність атак типу Prompt Leakage базується на експлуатації фундаментальної здатності моделей до повторення контексту, що