

А. В. Притула¹
Л. М. Куперштейн¹

МОДЕЛЬ ВІНАГОРОДИ В НАВЧАННІ З ПІДКРІПЛЕННЯМ ДЛЯ ЗАДАЧІ ТЕСТУВАННЯ БЕЗПЕКИ ВЕБЗАСТОСУНКІВ

¹Вінницький національний технічний університет

Досліджено задачу формування функції винагороди для навчання з підкріпленням в аналізі сценаріїв кібератак, у якій традиційні підходи в межах марковських процесів прийняття рішень, орієнтовані переважно на досягнення кінцевого стану, недостатньо відображають часову динаміку компрометації та неоднакову важливість проміжних цілей. Запропоновано модель винагороди, яка поєднує часову складову — час до компрометації (time-to-compromise, TTC) і семантичні ваги проміжних цілей, зокрема дій, пов'язаних із підвищенням привілеїв і доступом до критичних ресурсів. На відміну від традиційних схем винагороди, такий підхід формує локально інформативний сигнал підкріплення, що спрямовує навчання не лише на досягнення компрометації, а й на побудову часово ефективних і структурно змістовних траєкторій атаки. Експериментальне дослідження виконано на 12 стохастично згенерованих MAL-сценаріях (7200 запусків на метод), сформованих на основі відкритих моделей атак. Отримані результати показали, що запропонована модель забезпечує суттєво менше середнє значення TTC порівняно з трьома базовими підходами (перевага від 5 % до 21 %) з високим рівнем статистичної значущості, одночасно підвищуючи частоту успішного досягнення цільового стану і зменшуючи кількість недопустимих дій. Водночас побудовані політики стають інтерпретованішими: частка цільових дій у побудованих траєкторіях зростає, а допоміжні кроки витісняються на користь стратегічно значущих дій підвищення привілеїв. Аналіз чутливості підтвердив стійкість результатів до вибору параметрів моделі. Практичне значення роботи полягає у можливості використання запропонованого підходу для виявлення, ранжування та пріоритезації найнебезпечніших сценаріїв компрометації вебзастосунків з урахуванням їх часової та структурної критичності.

Ключові слова: марковський процес; кібербезпека; загроза; граф атаки; навчання з підкріпленням; нейронна мережа; функція винагороди; машинне навчання; штучний інтелект; моделювання.

Вступ

Моделювання кібератак засобами навчання з підкріпленням стало одним із найпомітніших напрямків автоматизації тестування на проникнення та побудови політик атак, оскільки воно дає змогу трактувати компрометацію як послідовний процес вибору дій у складному та частково структурованому просторі станів. Саме тому в останні роки активно розвиваються підходи, у яких навчання з підкріпленням використовується для пошуку траєкторій атак, автоматизації тестування на проникнення та роботи з графами атак, зокрема в роботах [1]—[4]. В україномовній літературі аналогічні підходи досліджуються для автоматизованого тестування на проникнення з використанням глибокого навчання з підкріпленням [5] та для виявлення вразливостей SQL-ін'єкцій [6]. Ці праці показують, що навчання з підкріпленням добре працює як інструмент послідовного пошуку траєкторій атак, але водночас висувають на перший план проблему коректного формування функції винагороди, оскільки саме вона визначає, чи буде агент оптимізувати справді небезпечну траєкторію, чи лише зручну з погляду формальної метрики.

Аналіз сучасних досліджень дозволяє виділити три основні напрямки [1]—[10], [11]—[16]:

- 1 — автоматизований пошук траєкторій атак засобами навчання з підкріпленням;
- 2 — використання метрики TTC як кількісної міри складності компрометації;
- 3 — вагове ранжування проміжних цілей на основі загальної системи оцінювання вразливостей (Common Vulnerability Scoring System, CVSS) або подібних оцінок.

Метрика TTC кількісно характеризує очікуваний час, необхідний атакувальнику для реалізації конкретного кроку атаки або повної траєкторії компрометації, та об'єднує складність знаходження



і реалізації експлойту з ймовірністю успіху у єдину величину з розмірністю часу. Система CVSS надає числові оцінки критичності вразливостей, що враховують технічні характеристики експлойту та потенційний вплив на систему. Обидві метрики відіграють ключову роль у формуванні функції винагороди для навчання з підкріпленням — TTC задає часову складову, а CVSS-подібні оцінки обґрунтовують семантичну вагу різних проміжних цілей.

У межах першого напрямку, який відповідає базовим схемам типу термінальної винагороди, сформувалася достатньо представницька група робіт, але їхні обмеження мають системний характер. У статті [1] навчання з підкріпленням застосовується для вибору дій у середовищі тестування на проникнення, однак оптимізація орієнтована передусім на досягнення успішного проходження сценарію, а не на деталізовану семантичну оцінку проміжних кроків. У роботі [2] проблема масштабування великого простору дій розв’язується через ієрархічну декомпозицію, але логіка винагороди залишається підпорядкованою переважно фінальному успіху. У дослідженні [3] наголос зроблено на компроміс між навчанням без моделі та апіорними знаннями, однак центральна складність проектування функції винагороди залишається відкритою. У статті [4] показано високу здатність глибокого навчання з підкріпленням до побудови траєкторій проникнення, але ефективність досягається насамперед через кращий пошук у просторі дій, а не через узгоджене моделювання часової та семантичної цінності дій. Отже, цей напрям демонструє силу навчання з підкріпленням як механізм пошуку послідовностей, але не дає задовільної відповіді на питання, якою саме повинна бути функція винагороди, щоб зробити такі послідовності одночасно швидкими й змістовно важливими.

У другому напрямку, який змістовно відповідає базовим схемам типу покрокового штрафування або оптимізації на основі TTC, центральним стає час до компрометації. У праці [5] закладено ранню основу для TTC як кількісної характеристики вразливості. У статті [6] запропоновано системне вимірювання середнього часу до компрометації на рівні мережі, що зробило час до компрометації інструментом кількісного порівняння станів захищеності. У роботі [7] TTC вже безпосередньо використовується в контексті аналізу графів атак. У статті [10] метрика TTC розширюється до оцінювання технік атак у контексті MITRE ATT&CK [11]. Усі ці праці важливі тим, що переводять ризик у часово вимірюваний простір. Проте саме в контексті навчання з підкріпленням вони залишають відкритим інше питання: чисто часовий штраф скорочує довжину або тривалість шляху, але не розрізняє, чи веде швидка дія до стратегічно важливої цілі, орієнтованої на підвищення привілеїв, чи лише до локального й менш критичного проміжного стану, тому TTC як основний компонент функції винагороди є потрібним.

Третій напрям, що відповідає зваженим базовим схемам з урахуванням критичності, намагається компенсувати цей недолік через семантичне ранжування цілей атаки. У роботі [12] закладено основу для числового представлення критичності вразливостей. У праці [13] CVSS-подібні оцінки інтегруються в орієнтований на навчання з підкріпленням аналіз графів атак. У роботі [14] функція винагороди пов’язується з критичністю дій у графі атак, доповненому навчанням з підкріпленням. У статтях [15]—[16] показано, що включення експертних знань або вагових коефіцієнтів на основі CVSS покращує спрямованість політик атак. Однак і тут зберігається обмеження: якщо функція винагороди надто сильно зміщується в бік критичності вузлів або вразливостей, то агент отримує стимул до семантично важливих, але не обов’язково часово-ефективних траєкторій. Відтак вагові схеми краще відрізняють важливі цілі атаки, але не розв’язують задачу балансування між важливістю проміжних цілей і часом до компрометації.

Проблема формування функції винагороди актуальна і в суміжних галузях, зокрема в задачах навігації автономних роботів [16].

Саме ця розірваність між підходами, орієнтованими на досягнення кінцевого стану, мінімізацію часу до компрометації та врахування критичності проміжних цілей, визначає наукове завдання, що розв’язується в межах цього дослідження, а саме побудову моделі винагороди, яка не зводить навчання агента лише до досягнення фінального стану, не обмежується часовою вартістю траєкторії, не покладається лише на вагове ранжування окремих кроків атаки, а забезпечує їх узгоджене врахування.

У попередньому дослідженні [18] запропоновано формалізацію сценаріїв кібератак у вигляді марковського процесу прийняття рішень із семантично обмеженим простором допустимих дій на MAL-графах та часовою інтерпретацією траєкторій через метрику часу до компрометації. Ця формалізація забезпечила коректне відтворення логіки проходження MAL-графа та усунення семантично нерелевантних переходів, проте питання структури функції винагороди в ній розглядалося лише на базовому рівні, як штраф за очікуваний час виконання кроку, без урахування відносної

семантичної важливості різних проміжних цілей. Унаслідок цього агент отримував однорідний сигнал підкріплення, який не розрізняв допоміжні кроки, кроки доступу до даних і кроки підвищення привілеїв, що обмежувало інтерпретованість отриманих політик. Саме це питання розв'язується в цьому дослідженні. Тут базова MDP-формалізація з [18] приймається як вихідна, а предметом дослідження є структура функції винагороди, що поєднує часову та семантичну складові.

Метою дослідження є підвищення ефективності та інтерпретованості політик атак шляхом удосконалення функції винагороди, що забезпечує зменшення часу до компрометації та підвищення частки семантично значущих дій у траєкторіях. Для досягнення мети необхідно проаналізувати наявні підходи до формування функції винагороди в задачах моделювання кібератак та виявити їх обмеження, на основі MDP-формалізації сценаріїв кібератак з [18] розробити модель винагороди, що одночасно враховує часову складову ТТС і семантичну важливість проміжних цілей, реалізувати навчання агента з використанням запропонованої функції винагороди, а також провести експериментальне дослідження й порівняльний аналіз різних функцій винагороди.

Наукова новизна роботи полягає в тому, що запропоновано модель винагороди, яка вперше поєднує часову вартість кроку атаки із семантичною вагою його класу в єдиному локально інформативному сигналі підкріплення. До того ж, введено показник інтерпретованості траєкторій як критерій оцінювання якості політики, що доповнює традиційні часові метрики. У цьому дослідженні, на відміну від [18], де формалізовано лише середовище навчання у вигляді MDP із семантично обмеженим простором дій, побудовано саму функцію винагороди та експериментально підтверджено її перевагу над трьома базовими схемами за сукупністю часових, структурних і поведінкових показників.

Формалізація середовища навчання

На відміну від класичних задач навчання з підкріпленням, де середовище має однорідну структуру дій, у моделюванні кібератак кроки принципово неоднорідні, оскільки, наприклад, розвідувальні дії, підвищення привілеїв і доступ до даних мають різну стратегічну вагу і різну часову вартість. Ця неоднорідність і визначає вимоги до структури функції винагороди.

Побудова моделі винагороди потребує формального середовища, в якому визначено простір станів, множину допустимих дій і механізм переходів. Як таке середовище використовується MDP-формалізація сценаріїв кібератак на MAL-графах, запропонована в [18]. У межах цієї формалізації визначено простір станів, станозалежну множину допустимих дій з урахуванням AND/OR-залежностей між кроками атаки та базову часову інтерпретацію траєкторій через метрику часу до компрометації.

У класичних постановках навчання з підкріпленням для задач планування атаки функція винагороди концентрується переважно на досягненні термінального стану. Така схема не розрізняє траєкторій, що ведуть до однакової цілі за істотно різний час, і не враховує того, що проміжні стани можуть мати різну безпекову вагу. Наприклад, компрометація сервісного облікового запису з низькими привілеями і компрометація адміністратора не є рівнозначними навіть у тому випадку, коли жодна з цих дій ще не є фінальним станом атаки. Відповідно, для задач моделювання кібератак доцільним є перехід до такої функції винагороди, у якій принципово важливими є не лише сам факт компрометації, а й час її досягнення та семантична важливість проміжних кроків.

Середовище навчання задається множиною MAL-графів атак [19], що стохастично генеруються з ієрархії класів специфікації MAL. Навчання виконується не на одному фіксованому графі, а на розподілі середовищ $G \sim \mathcal{G}$, тому оптимальна політика має узагальнюватися на клас екземплярів, а не адаптуватися до однієї конкретної топології. Відповідно, задача оптимізації політики визначається як

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{G \sim \mathcal{G}} \left[\mathbb{E}_{\pi} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t, s_{t+1}) \right] \right], \quad (1)$$

де π — політика агента (відображення станів у розподіл дій); $\mathbb{E}$ — математичне сподівання; G — випадковий граф атаки з розподілу середовищ; $\mathcal{G}$ — множина (розподіл) можливих графів атак; t — дискретний крок часу; T — довжина епізоду; $\gamma \in [0,1]$ — коефіцієнт дисконтування; s_t — стан системи на кроці t ; a_t — дія, виконана на кроці t ; s_{t+1} — наступний стан після виконання дії; $r(\cdot)$ — функція винагороди. Для задачі моделювання кібератак це означає, що саме форма $r(\cdot)$ визначає, чи буде агент орієнтований на швидке досягнення критичних вузлів, чи лише на формальне досягнення деякого фінального стану без урахування структури траєкторії.

Генерація середовища базується на стохастичній генерації кількості екземплярів класів MAL. Для кожної групи компонентів, що відповідає певному типу об'єктів системи, число екземплярів визначається через біноміальний розподіл

$$N_i \sim \text{Bi}\left(\frac{\mu_i^2}{\mu_i - \sigma_i^2}, 1 - \frac{\sigma_i^2}{\mu_i}\right), \quad (2)$$

де μ_i і σ_i^2 є відповідно середнім значенням та дисперсією числа екземплярів класу i . Параметризація є коректною за умови $\sigma_i^2 < \mu_i$, що забезпечує $p \in (0,1)$. Така параметризація забезпечує контроль над масштабом системи та її структурною варіативністю. У термінах кібербезпеки це дозволяє моделювати як компактні вебсередовища, що складаються з одного застосунку, однієї бази даних і обмеженої кількості облікових записів, так і складні конфігурації з кількома сервісами, різними рівнями доступу та множиною критичних цілей.

Кожний вузол згенерованого графа атаки задається як семантична сутність [18]

$$v = (c, i, \alpha), \quad (3)$$

де c є класом активу MAL-моделі (наприклад, *Application, Network, Identity*), i є індексом конкретного екземпляра цього класу в досліджуваній системі, a — кроком атаки, визначеним у межах класу c . Наприклад, для вузла *Application.Service₂.network RequestConnect*, $c = \text{Application}$ (клас активу), $i = \text{Service}_2$ (другий екземпляр цього класу в системі), $\alpha = \text{networkRequestConnect}$ (спроба встановлення мережевого з'єднання), тобто описує спробу встановлення з'єднання із сервісом під номером 2, в той час, як *Identity.Admin₁.compromise* описує компрометацію користувача з правами адміністратора. Таке подання дає можливість відрізнити кроки типу *networkRequestConnect*, *SuccessfulAccess*, *readData*, *writeData* або *credentialAccess* не лише структурно, а й функціонально. Саме семантична структура (3) є основою для подальшого визначення функції ваг $w(a)$, оскільки дозволяє розрізнити допоміжні, цільові та критичні кроки атаки.

Кожному кроку атаки a ставиться у відповідність невід'ємна випадкова величина $ttc(a) \sim D_a, ttc(a) \geq 0$, де D_a — розподіл часу для виконання цієї дії, а сама величина інтерпретується як час до компрометації. У кібербезпековій інтерпретації це означає, що, наприклад, SQL-ін'єкція, компрометація сесії, горизонтальне переміщення в мережі між сервісами або підвищення привілеїв можуть мати різні часові характеристики. Саме ця різниця робить недоцільним використання схем винагороди, орієнтованих лише на фінальний стан, оскільки дві траєкторії з однаковою кінцевою ціллю можуть істотно відрізнитися за часом досягнення критичного ресурсу.

Граф атаки $G = (V, E)$ інтерпретується як середовище марковського процесу прийняття рішень $M = (S, A, P, R, \gamma)$ у постановці, прийнятій в [18]: S — множина станів, A — множина дій, $P(s' | s, a)$ — ймовірність переходу зі стану s у стан s' при дії a , $R(s, a, s')$ — функція винагороди, γ — коефіцієнт дисконтування.

У цій постановці стан агента визначається не повним графом, а підграфом поточної досяжності $F_t = (V_t, E_t)$, де множина вершин стану задається співвідношенням

$$V_t = U_t \cup \{v \in N(v') : v' \in U_t\}, \quad (4)$$

де U_t є множиною вже виконаних кроків атаки, тоді як $N(v)$ позначає множину нащадків вершини v . Таким чином, стан містить як виконані кроки, так і ті дії, що стають доступними для негайного виконання. У прикладному сенсі це відповідає поведінці атакувальника, який після компрометації певного сервісу відкриває для себе новий локальний простір можливостей, але не має повного контролю над усім графом атаки. Саме це подання робить винагороду локально інформативною, оскільки агент оцінює не абстрактні глобальні дії, а ті кроки, що реально доступні в поточній конфігурації системи.

Базовий простір дій визначається як множина доступних переходів із поточного підграфа

$$A(s_t) = \{v_j | \exists v_i \in U_t : (v_i, v_j) \in E\}. \quad (5)$$

У такому вигляді враховується лише структурна досяжність, тоді як у MAL істотну роль відіграють кроки типу AND, для яких виконання можливе лише за умови одночасного виконання всіх передумов. Тому у [18] базове визначення простору дій уточнюється обмеженням допустимості, за яким для кожної дії a необхідне виконання всіх її батьківських кроків

$$Pre(a) \subseteq U_t, \quad (6)$$

де $Pre(a)$ — множина батьківських кроків атаки для дії a . Це обмеження є принциповим і для задачі формування функції винагороди, адже без нього винагорода нараховувалася б за структурно видимі, але семантично нереалізовані дії, що генерувало б корисні сигнали для нефізичних траєкторій. У контексті кібербезпеки наведене обмеження означає, що, наприклад, доступ до захищеної бази даних може бути дозволений лише після компрометації вебсервера та отримання відповідних облікових даних. Відповідно, перехід системи в новий стан виконується виключно на допустимих діях $A(s_t)$ і відповідає приєднанню виконаної дії до множини U_t виконаних кроків, що визначає коректне оновлення підграфа поточної досяжності з (4).

Модель винагороди

Після того як прийнято опис стану, простору дій (5) і обмеження допустимості (6), відкритим лишається центральне питання — як саме задавати функцію винагороди R на множині допустимих дій. У роботі [18] для цього використовувався штраф за очікуваний час виконання кроку, який забезпечує часову оптимізацію траєкторій, але не розрізняє допоміжні дії, дії доступу до даних і дії підвищення привілеїв. Саме цю прогалину і закриває схема винагороди, що будується нижче.

Пропонується ввести схему винагороди, у якій значення r для дії a визначається одночасно часовою складовою, що відображає очікувану вартість виконання цієї дії, та семантичною вагою, що відображає її критичність для атаки:

$$r(a) = w(a) - \lambda \cdot ttc(a), \quad (7)$$

де $\lambda > 0$ є коефіцієнтом масштабування часової складової; $w(a)$ — семантична вага кроку. Оскільки в задачі компрометації семантична вага та ТТС є властивостями самої дії, а не пари станів, функція винагороди спрощується до $r(a)$. Центральна ідея моделі полягає в тому, що агент не просто уникає тривалих дій, а оцінює кожну дію з погляду того, наскільки її цінність для атаки компенсує часові витрати на її виконання. Якщо дія є швидкою, але не веде до стратегічно важливого результату, її внесок у сумарну винагороду лишається малим. Якщо ж дія відкриває шлях до привілейованого доступу або критичних даних, вона може бути доцільною навіть за більшого $ttc(a)$.

Семантична вага визначається через належність дії до одного з класів цілей

$$w(a) = \begin{cases} w_p, a \in A_{privilege}, \\ w_d, a \in A_{data}, \\ 0, a \in A_{aux}, \end{cases} \quad w_p > w_d > 0. \quad (8)$$

У запропонованій моделі прийнято значення $w_p = 2$ для кроків підвищення привілеїв, $w_d = 1$ для кроків, безпосередньо пов'язаних із компрометацією даних, та нульова вага для допоміжних кроків. Таке співвідношення фіксує ієрархію важливості подій у вебзастосунку, у якій адміністративний контроль над системою розглядається як приблизно вдвічі критичніший за локальний доступ до окремих даних, а допоміжні дії (наприклад, сканування або проміжні мережеві з'єднання) не повинні штучно стимулюватися і отримують нульову вагу. Співвідношення 2:1 для привілейованих і даних-орієнтованих кроків узгоджується з пріоритетами, прийнятими у відкритих моделях загроз типу MITRE ATT&CK [11], де підвищення привілеїв розглядається як критичний етап, що відкриває множину подальших траєкторій атаки. Значення $\lambda = 0,5$ обрано за результатами аналізу чутливості (який описано далі), у якому встановлено плоску область мінімуму ТТС для $\lambda \in [0,5, 1,0]$. У цьому діапазоні часовий штраф і семантичний бонус мають зіставні порядки величини, що забезпечує баланс між мінімізацією часу компрометації та керуванням агента до стратегічно важливих кроків.

Для усунення залежності від абсолютного масштабу $ttc(a)$ вводиться також нормована форма винагороди

$$\tilde{r}(a) = \frac{w(a)}{w_p} - \lambda \frac{ttc(a)}{T_{max}}, \quad (9)$$

де T_{max} верхньою є оцінкою ТТС у поточному середовищі. Нормування не змінює концептуальної природи моделі, але підвищує стійкість навчання на множині різнорозмірних графів, оскільки часові характеристики окремих кроків можуть мати різний порядок величини залежно від типу системи та складності конкретного експлоїту. У подальшому використовується саме нормована форма $\tilde{r}(a)$, як така, що забезпечує узгоджену поведінку агента на графах з різною кількістю вузлів і різними масштабами ТТС.

Із підстановки (7) у стандартний критерій дисконтованого повернення випливає, що максимізація очікуваної сумарної винагороди агента еквівалентна мінімізації зваженої суми двох складових, очікуваного часу виконання кроків атаки і від'ємного вкладу семантичної ваги вибраних дій

$$\min_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^T \gamma^t ttc(a_t) - \frac{1}{\lambda} \sum_{t=0}^T \gamma^t w(a_t) \right], \quad (10)$$

у такій формі цільова функція набуває прямої інтерпретації. Політика оптимізує не просто досягнення фінального стану, а компроміс між часовою ефективністю траєкторії та її стратегічною значущістю. Агент навчається вибирати не будь-який шлях до компрометації, а той, який мінімізує час до критичних ресурсів і водночас проходить через найзначущіші для атаки вузли.

Для термінального вузла v_t приймається $r(v_t) = 0$. Нульова винагорода не стимулює передчасного завершення епізоду, оскільки відмова від подальших дій є вигідною лише тоді, коли всі доступні продовження мають настільки високий часовий штраф або настільки низьку семантичну цінність, що продовження траєкторії стає нераціональним. Тобто термінальний вузол виконує роль механізму, який дозволяє політиці балансувати між продовженням атаки та зупиненням, а не лише формально фіксувати завершення епізоду.

Для оптимізації політики в межах описаного MDP-середовища застосовується стандартний алгоритм Deep Q-Network [20] у тій самій архітектурній конфігурації, що й у [18]. Вибір DQN зумовлений експоненційною розмірністю простору станів і стандартними стабілізаційними механізмами — буфером досвіду і цільовою мережею. Запропонована функція винагороди у нормованій формі (9) безпосередньо входить у формування цільового значення для оновлення Q-функції. На кожному кроці агент отримує миттєву винагороду $\tilde{r}(a_t)$ за виконану дію, а максимум по наступному стану береться виключно на множині семантично допустимих дій $A(s_{t+1})$, визначеній обмеженням (6). Тому зміна структури винагороди змінює не лише інтерпретацію оптимальної політики, а й саму динаміку навчання агента.

Отже, запропонована модель винагороди ґрунтується на трьох взаємопов'язаних принципах: ТТС-орієнтованість через часовий штраф, семантична зваженість через ваги проміжних цілей, визначеність лише на семантично допустимих діях. Поєднання цих принципів у єдиному локально інформативному сигналі підкріплення спрямовує агента на побудову часово ефективних і структурно інтерпретованих сценаріїв атаки.

Експерименти та результати

Експериментальна частина спрямована на перевірку того, чи змінює запропонована модель винагороди не лише фінальні значення метрик, а й сам режим навчання агента, структуру побудованих траєкторій та ступінь їх інтерпретованості в задачах моделювання кібератак. Результати аналізуються одночасно в трьох взаємопов'язаних площинах, а саме в площині динаміки навчання, у площині структурних характеристик сформованих шляхів та в площині чутливості до параметрів функції винагороди.

Усі результати отримано на множині з дванадцяти сценаріїв, сформованих не довільно, а шляхом генерації MAL-графів на основі відкритих моделей атак і відтворюваного стохастичного процесу генерації. Формальну семантику кроків атаки та їх залежностей взято з enterpriseLang [21], який реалізує доменно-специфічну мову MAL і містить типові шаблони компрометації, зокрема експлуатація загальнодоступних застосунків, підвищення привілеїв та горизонтальне переміщення в мережі. Для прив'язки цих шаблонів до веб-орієнтованого середовища прикладні вузли та їх послідовності зіставлено з відкритими навчальними сценаріями OWASP WebGoat [22], що охоплюють

SQL-ін'єкцію, обхід автентифікації, небезпечне пряме посилання на об'єкт та інші характерні кроки атак. Формування конкретних графів здійснювалося шляхом стохастичного варіювання структури з незалежною зміною кількості екземплярів вузлів, глибини графа, частки AND- і OR-залежностей та конфігурації міжвузлових зв'язків. Часові характеристики окремих кроків атаки і їх варіативність задавалися через емпірично узгоджені розподіли ТТС, що відтворюють реалістичну складність веб-орієнтованих атак. Загалом згенеровано 12 сценаріїв як компроміс між різноманітністю структур і обчислювальною складністю експериментів. Порівняно з попередньою роботою [18] кількість збільшено для повнішого аналізу чутливості до параметрів функції винагороди. У результаті кожен сценарій є стохастично згенерованим MAL-графом з параметрами $|V| \in [28, 48]$ (кількість вершин графа), $|E| \in [45, 90]$ (кількість ребер графа), глибиною 3...10, кількістю AND-вузлів 8...20 і OR-вузлів 4...9. Діапазони розширено як для коротких атак з глибиною 3, так і глибоких багатокрокових траєкторій. Ці параметри варіюються незалежно, тому складність сценаріїв не є монотонною функцією розміру графа. Структурні характеристики кожного сценарію подано в табл. 1.

Таблиця 1

Структурні характеристики сценаріїв кібератак

№ сценарію	Вузли	Ребра	AND-вузли	OR-вузли	Базові вузли	Середнє ТТС кроку	Глибина
1	28	45	9	4	15	2.9	3
2	34	58	11	5	18	3.2	4
3	41	72	16	6	19	3.8	8
4	30	51	10	8	12	3.1	5
5	47	87	18	7	22	4.0	9
6	36	63	13	5	18	3.5	6
7	29	54	8	9	12	3.0	4
8	43	78	15	8	20	3.7	7
9	32	56	10	6	16	3.3	5
10	45	82	17	6	22	3.9	7
11	48	90	20	5	23	4.2	10
12	37	66	14	7	16	3.6	6

Наведене значення ТТС відповідає середньому ТТС одного кроку атаки. Сценарії утворюють неоднорідну множину з різними топологічними та семантичними характеристиками. Зокрема, сценарії 3, 5 та 11 характеризуються більшою глибиною та домінуванням AND-залежностей, що звужує множину допустимих дій і ускладнює процес навчання. Натомість сценарії 4 і 7 мають вищу частку OR-вузлів, що формує розгалуженіший простір альтернативних траєкторій.

Для порівняльного аналізу вибрано три базових методи, що репрезентують найпоширеніші схеми формування функції винагороди в задачах навчання з підкріпленням. Метод термінальної винагороди (далі — TerminalReward) нараховує ненульову винагороду лише у момент досягнення цільового вузла, тоді як на всіх проміжних кроках вона дорівнює нулю [1]. Метод покрокового штрафу (далі — StepPenalty) доповнює термінальну винагороду фіксованим негативним штрафом на кожному кроці, що стимулює скорочення довжини траєкторії [3]. Метод статичної винагороди (далі — StaticReward), навпаки, нараховує фіксовану позитивну винагороду на кожному проміжному кроці незалежно від семантики виконаної дії та її часової вартості [4]. Спільна риса всіх трьох підходів полягає в тому, що вони не враховують семантику виконаної дії під час формування винагороди. Усі методи, включно із запропонованим (далі — Proposed), діють в однаковому середовищі з єдиною множиною допустимих дій $A(s)$, визначеною формулою (6). Відмінність полягає виключно у структурі функції винагороди.

Для кожного сценарію згенеровано 15 різних графів, і для кожного виконано 40 незалежних запусків кожного методу. Отже, експериментальна множина становить 7200 запусків на метод і 28800 запусків у сукупності, що дозволяє розглядати отримані оцінки як статистично репрезентативні на рівні сценаріїв, графів та окремих запусків.

Усі методи реалізовані на одній DQN-архітектурі, що дозволяє коректно порівнювати результати. Для апроксимації Q-функції використовується повнозв'язна нейронна мережа з двома прихованими шарами розмірності 128 і 64 нейрони з активацією ReLU. Як оптимізатор застосовано Adam з коефіцієнтом навчання 10^{-3} . Розмір буфера досвіду становить 50000 переходів, мінібатч для оновлення параметрів — 64 зразки. Цільова мережа оновлюється кожні 1000 кроків навчання. Для балансу між дослідженням і експлуатацією використовується ϵ -жадібна стратегія з початковим

значенням $\varepsilon = 1,0$, яке лінійно зменшується до $\varepsilon = 0,05$ протягом перших 60 % епізодів навчання і далі фіксується. Навчання триває 200 епізодів для кожного графу, коефіцієнт дисконтування $\gamma = 0,95$.

Вхід до Q-мережі формується для кожної допустимої дії $a \in A(s)$ окремо, тобто апроксимація виконується для функції $Q(s, a)$, а не у вигляді вектора значень для всіх дій одночасно. Для побудови подання стану s використовується графовий модуль уваги GAT, застосований до підграфа поточної досяжності V_t , який відповідає стану агента в термінах формули (4). GAT-модуль складається з двох послідовних шарів багатоголової уваги з чотирма головами на кожному шарі. Вихідна розмірність вузлового ембедінгу після другого шару становить 128. На основі множини вузлових ембедінгів підграфа V_t обчислюється глобальне представлення стану $h_G \in \mathbb{R}^{128}$ за допомогою операції global mean pooling. Для кожної допустимої дії $a = (u, v)$, де $u \in U_t$ — вже компрометований вузол, а v — цільовий вузол атаки відповідно до (5), додатково використовуються ембедінг вихідного вузла $h_u \in \mathbb{R}^{128}$, ембедінг цільового вузла $h_v \in \mathbb{R}^{128}$, скалярне значення $TTC(a)$ для цієї дії та скаляр, що задає максимальну довжину епізоду L_{max} . У результаті вхідний вектор Q-мережі визначається як конкатенація $x(s, a) = [h_G \| h_u \| h_v \| TTC(a) \| L_{max}]$ і має розмірність $128 + 128 + 128 + 1 + 1 = 386$. Q-мережа є повнозв'язною і для кожної пари (s, a) повертає одне скалярне значення $Q(s, a)$, яке інтерпретується як оцінка очікуваного дисконтованого повернення у разі вибору дії a у стані s . На етапі вибору дії використовується ε -жадібна стратегія: з ймовірністю $1 - \varepsilon$ вибирається дія з максимальним значенням $Q(s, a)$ серед усіх допустимих дій $A(s)$, а з ймовірністю ε виконується випадковий вибір з тієї самої множини (рис. 1). Реалізацію виконано мовою Python з використанням бібліотек PyTorch, PyTorch Geometric для графового модуля GAT, Gymnasium для моделювання MDP-середовища, TorchRL для реалізації DQN-агента та NetworkX для роботи зі структурою графів атаки.

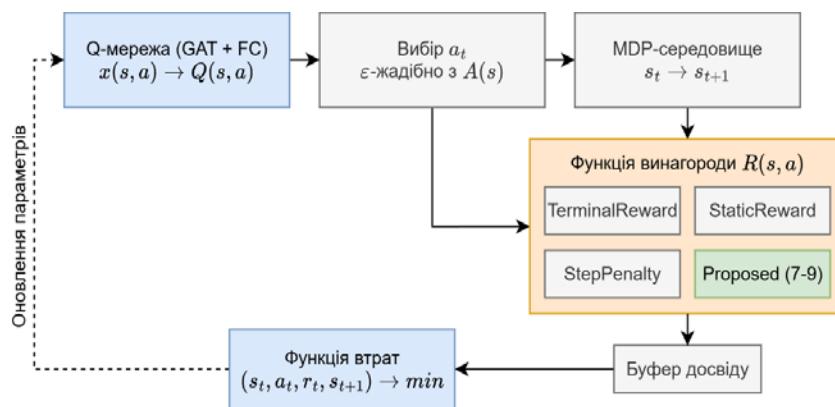


Рис. 1. Схема циклу навчання DQN

Динаміку навчання та фінальні розподіли показників для всіх методів показано на рис. 2.

Схожість форм кривих на рис. 2а і 2б пояснюється тим, що накопичена винагорода за формулою (7) лінійно залежить від TTC, тому ці дві величини математично пов'язані. Однаковий характер динаміки зумовлений спільною DQN-архітектурою та ε -стратегією. Різниця між методами проявляється не у формі кривих, а в асимптотичних значеннях, до яких вони збігаються, що підтверджується числовими результатами, поданими в табл. 2.

Уже на нульовому епізоді середній TTC для Proposed є нижчим ніж для всіх трьох базових підходів, і ця перевага зберігається протягом усього навчання. До 200-го епізоду Proposed досягає середнього значення TTC, що відповідає підсумковим значенням у табл. 2, з перевагою близько 5 % над найсильнішим базовим методом StepPenalty і понад 20 % над TerminalReward.

Криві на рис. 2а не є штучно гладкими. Поблизу 60-го, 110-го і 160-го епізодів спостерігаються локальні відхилення, що відповідає стохастичному характеру процесу навчання. Динаміка накопиченої винагороди на рис. 2б демонструє ту саму закономірність — схема винагороди з часовою та семантичною складовими не лише скорочує TTC, а й забезпечує щільніший і стабільніший сигнал підкріплення. З рис. 2в видно, що медіана епізоду збіжності для Proposed є найменшою, причому розподіл лежить лівіше від усіх базових методів, агент не просто приходить до кращого фінального результату, а робить це швидше. На рис. 2г медіана і міжквартильний розмах фінального TTC для Proposed зміщені в область менших значень, однак розподіл частково перекривається з StepPenalty.

Це перекриття показує, що перевага Proposed не є механічно гарантованою для кожного окремого графа, а зумовлена статистичною перевагою на множині сценаріїв.

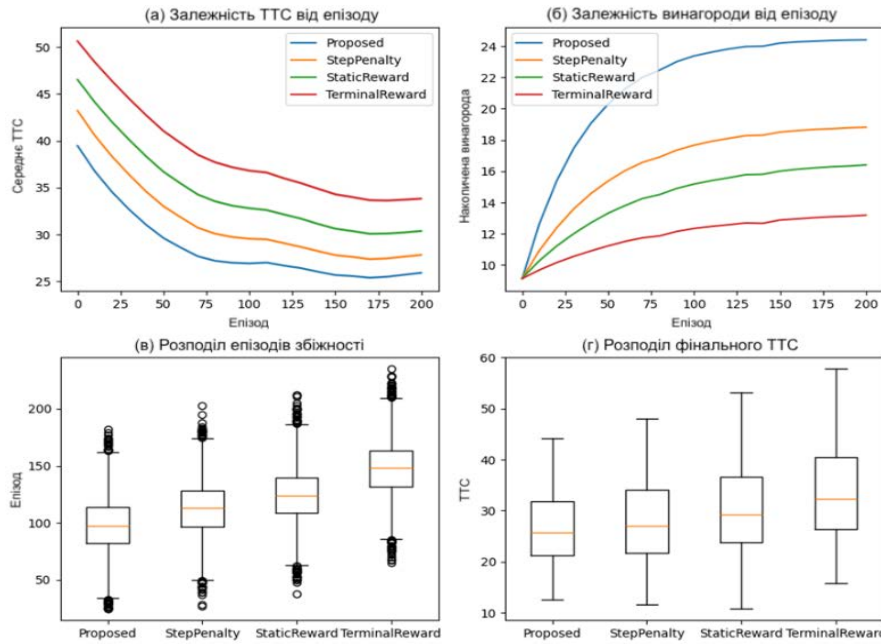


Рис. 2. Динаміка навчання та фінальні розподіли показників для досліджуваних методів: а — залежність TTC від епізоду; б — залежність винагороди від епізоду; в — розподіл епізодів збіжності; г — розподіл фінального TTC)

Таблиця 2

Агреговані результати порівняння методів за часовими, структурними та статистичними характеристиками

Метрика	Proposed	StepPenalty	StaticReward	TerminalReward
Середнє TTC траєкторії	26,637	28,002	30,382	33,562
Стандартне відхилення TTC	6,581	7,379	7,818	8,546
95 % ДІ TTC нижня межа	26,547	27,904	30,280	33,447
95 % ДІ TTC верхня межа	26,727	28,100	30,484	33,677
р-значення порівняно з Proposed	—	< 1e-6	< 1e-6	< 1e-6
Середня довжина траєкторії	7,470	7,619	8,441	9,471
Стандартне відхилення довжини траєкторії	2,161	2,278	2,492	2,876
95 % ДІ довжини траєкторії (нижн.)	7,440	7,586	8,407	9,430
95 % ДІ довжини траєкторії (верхн.)	7,500	7,652	8,475	9,512
Середня накопичена винагорода	7,726	6,224	5,602	4,362
Стандартне відхилення накопиченої винагороди	4,107	3,943	3,781	3,522
Частота успішного завершення	0,8274	0,7989	0,7439	0,6872
Середня кількість нерелевантних дій	1,613	2,416	3,569	5,059
Загальна кількість запусків на метод	7200			
Загальна кількість графів	180			
Показник ефективності (11)	1,517	1,394	1,196	1,000

Новизна запропонованого підходу пов'язана не лише зі скороченням часу до компрометації, а й з отриманням політик, інтерпретованих у термінах кібербезпеки, де вибір дій зміщується до семантично значущих проміжних цілей.

З рис. 3а—в випливає, що запропонована модель справді змінює структуру поведінки агента. Частка допоміжних кроків для Proposed є найменшою серед усіх методів. Proposed витрачає меншу частину траєкторії на допоміжні дії й швидше переходить до цільових вузлів. Водночас частка привілейованих кроків для Proposed є найбільшою, що є суттєвим зсувом, оскільки саме ці дії відкривають найнебезпечніші шляхи подальшої компрометації. Розподіли частки цільових дій на рис. 3б підтверджують цю закономірність. Proposed формує траєкторії з більшою часткою стратегічно релевантних дій, ніж усі базові методи, причому розподіли для Proposed, StepPenalty і StaticReward частково перекриваються — тобто перевага не є штучно ідеалізованою. Аналогічну картину дає інтегральний показник інтерпретованості на рис. 3в, де перевага Proposed над StepPenalty становить

близько 13 %, над StaticReward — близько 19 %, а над TerminalReward — понад 43 %. Отже, запропонована схема винагороди не лише скорочує час до компрометації, а й систематично зміщує політику в бік змістовніших для аналізу безпеки траєкторій.

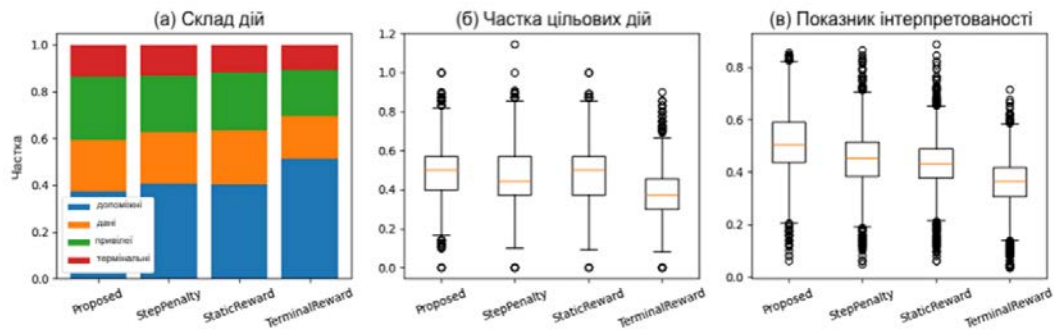


Рис. 3. Структурні характеристики сформованих траєкторій атак для різних методів:
а — склад дій; б — частка цільових дій; в — показник інтерпретованості

Окремої перевірки потребує питання стійкості результатів до параметрів самої функції винагороди. Якщо перевага моделі виникає лише в одній точці параметричного простору, то така перевага має обмежену методичну цінність (рис. 4).

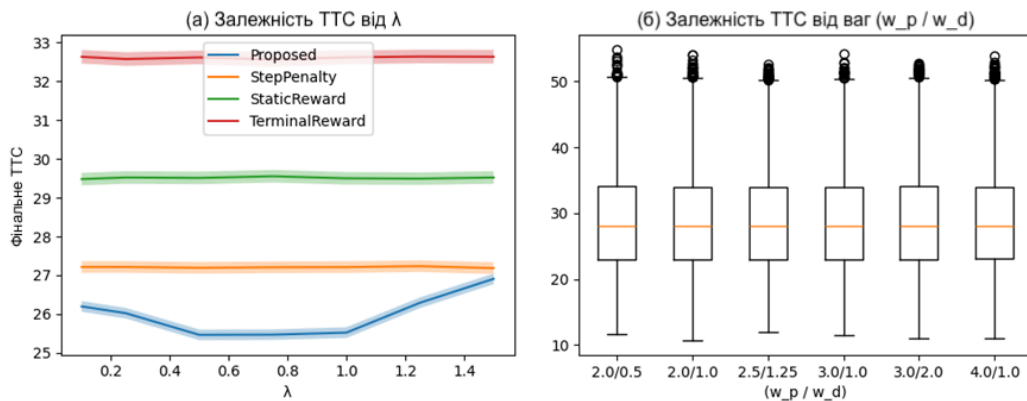


Рис. 4. Чутливість результатів до параметрів функції винагороди: а — залежність TTC від λ ;
б — залежність TTC від ваг (w_p, w_d)

Аналіз рис. 4 показує, що залежність результатів від параметрів функції винагороди є виразною, але не ідеально гладкою. Для запропонованої моделі мінімальні значення TTC спостерігаються в області $\lambda \in [0,5, 1,0]$, причому у цьому діапазоні відмінності між сусідніми значеннями λ лежать у межах ширини 95 % довірчого інтервалу. Це означає, що мінімум є не вузькою точкою, а плоскою областю параметричної стійкості. Поза цим діапазоном, коли $\lambda = 0,10$ і $\lambda = 0,50$ TTC помітно зростає, тобто часовий штраф не є нейтральним параметром, проте і надтонкого ручного налаштування не вимагає. Крива на рис. 4а містить локальні реверси та асиметрію, а довірчі інтервали не зводяться до нуля, що відповідає реалістичній стохастичній природі середовища. Аналіз рис. 4б показує, що найкращі значення TTC для Proposed одержано при парах (w_p, w_d) у діапазоні від 2:1 до 2.5:1.25. При зміщенні ваг у бік надмірного пріоритету привілейованих кроків (наприклад, 4:1) TTC зростає, а середня довжина шляху також збільшується. Це означає, що функція винагороди не реагує на зростання w_p монотонним покращенням. Надмірний перекид у бік дій підвищення привілеїв може змушувати політику обирати семантично сильні, але часово дорожчі траєкторії.

В таблиці агрегованих метрик (табл. 2) подано узагальнені результати, що дозволяють перевірити, чи повторюються висновки за різними показниками одночасно, а не лише за окремими візуальними трендами. Для цього потрібно зіставити TTC, довжину траєкторії, інтегральну ефективність, накопичену винагороду, частоту успіху та кількість недопустимих дій.

Для перевірки статистичної значущості різниці між методом Proposed і базовими підходами використовується двовибірковий t-тест Велча, який не передбачає припущення про рівність дисперсій. Такий вибір обґрунтований тим, що вибірки TTC для різних методів формуються незалежно на

стохастично згенерованих MAL-графах і мають помітно відмінні дисперсії. Для кожного методу формується вибірка значень TTC з усіх незалежних запусків (12 сценаріїв * 15 графів * 40 запусків), після чого виконується попарне порівняння Proposed з кожним базовим методом. Нульова гіпотеза — відсутність різниці середніх значень TTC. Різниця вважається статистично значущою з $p < 0,05$.

Отримані p -значення, наведені в табл. 2, задовольняють нерівність $p < 10^{-6}$, що підтверджує значущість спостережуваних різниць. Основною часовою метрикою в табл. 2 є сумарний TTC траєкторії — сума TTC усіх кроків побудованого шляху, усереднена по незалежних запусках. За порядком величини він відповідає добутку середнього TTC кроку (табл. 1) на середню довжину траєкторії.

З табл. 2 випливає, що Proposed забезпечує найменше середнє значення TTC при найменшому стандартному відхиленні. Перевага над StepPenalty становить близько 5 %, над StaticReward — близько 12 %, над TerminalReward — близько 21 %. Водночас довжина траєкторії для Proposed лише незначно менша за StepPenalty (7,47 проти 7,62), але істотно менша за StaticReward і TerminalReward. Це означає, що основна перевага Proposed полягає не в радикальному скороченні числа кроків порівняно з найсильнішим базовим методом, а в кращому поєднанні довжини шляху та його часової вартості.

Частота успішних проходжень є найбільшою саме для Proposed, а кількість недопустимих дій є найменшою. Таким чином, запропонована модель не є найкращою лише за одним окремим критерієм, а формує збалансованіший результат одразу за кількома часовими, структурними й статистичними характеристиками.

Для кількісної оцінки ефективності методів розраховується показник відносної ефективності, що порівнює кожен метод з TerminalReward за часом до компрометації і частотою успішного завершення одночасно

$$E = \frac{\overline{TTC}_{base}}{\overline{TTC}_{method}} \cdot \frac{P_{success, method}}{P_{success, base}}, \quad (11)$$

де $\overline{TTC}_{base}$, $P_{success, base}$ — відповідні значення середнього TTC та часток успішного завершення для TerminalReward як базової лінії, $\overline{TTC}_{method}$, $P_{success, method}$ — значення для порівнюваного методу. Перший множник показує, у скільки разів метод швидший за базовий, другий — у скільки разів він надійніший, а їх добуток характеризує сумарну перевагу. Для Proposed $E = 1,517$, що означає приблизно півторакратне перевищення базової лінії одночасно за часом і надійністю. Для StepPenalty $E = 1,394$, для StaticReward — 1,196. Proposed зберігає лідерство і в цьому агрегованому вимірі, але важливіше те, що він домінує над трьома базовими методами: у табл. 2 жоден з базових методів не має кращого значення за жодним з чотирьох показників — часом до компрометації, частотою успішного завершення, середньою накопиченою винагородою і кількістю недопустимих дій.

Оскільки новизна запропонованої моделі прямо пов'язана з інтерпретованістю політики, потрібне окреме узагальнення поведінкових характеристик агента (табл. 3).

Таблиця 3

Показники інтерпретованості траєкторій атаки для досліджуваних методів

Метрика	Proposed	StepPenalty	StaticReward	TerminalReward
Частка допоміжних дій	0,3739	0,4068	0,4043	0,5136
Частка дій з доступу до даних	0,2199	0,2237	0,2278	0,1783
Частка дій підвищення привілеїв	0,2723	0,2388	0,2490	0,2012
Частка термінальних дій	0,1339	0,1307	0,1189	0,1069
Середня кількість критичних кроків на траєкторію	3,677	3,517	4,032	3,608
Середня кількість допоміжних кроків на траєкторію	2,792	3,100	3,411	4,865
Частка цільових дій	0,4925	0,4625	0,4778	0,3807
Відношення критичних дій до допоміжних	1,7746	1,5116	1,5727	0,9119
Показник інтерпретованості	0,5151	0,4543	0,4323	0,3592
Загальна кількість запусків на метод	7200			
Загальна кількість графів	180			

Табл. 3 демонструє, що Proposed має найнижчу частку допоміжних дій і найвищу частку дій підвищення привілеїв серед усіх методів. Це означає, що модель зміщує політику не просто до довільних цільових кроків, а саме до критичніших дій, пов'язаних з привілеями. Частка цільових дій і відношення критичних до допоміжних також найбільші для Proposed, а інтегральний показник

інтерпретованості перевищує найсильніший базовий метод StepPenalty на 13 %. Отже, запропонована модель не лише знижує ТТС, а й систематично спрямовує політику до змістовніших, з погляду аналізу безпеки, траєкторій, у яких частка стратегічно важливих дій більша, а навантаження допоміжними кроками менше.

Висновки

У роботі запропоновано модель винагороди для задач навчання з підкріпленням у моделюванні кібератак, в якій інтегрується в єдину схему часову складову компрометації ТТС і семантичну важливість проміжних цілей (через ваги дій підвищення привілеїв і доступу до даних). На відміну від традиційних схем винагороди, орієнтованих переважно на досягнення кінцевого стану, запропонований підхід надає агенту локально інформативний сигнал підкріплення, що враховує не лише факт компрометації, а й часову вартість дії та її критичність у структурі атаки. Це забезпечило перехід від оптимізації самого факту компрометації до оптимізації часово ефективної та змістовно інтерпретованої траєкторії атаки.

Експериментальне дослідження на 12 стохастично згенерованих MAL-сценаріях підтвердило перевагу запропонованої моделі за всіма ключовими показниками: середнім ТТС (перевага 5—21 % над базовими підходами, $p < 10^{-6}$), частотою успішного завершення атаки, кількістю недопустимих дій та показником інтерпретованості траєкторій. Модель не лише скорочує час до компрометації, а й зміщує політику агента до семантично значущих дій, зокрема дій підвищення привілеїв, що підтверджується найвищим показником відносної ефективності (1,52 проти 1,00 для базової лінії).

Практичне значення полягає у можливості використання моделі для пріоритизації захисних заходів у вебзастосунках не лише за фактом існування траєкторії атаки, а й за швидкістю досягнення критичних ресурсів та щільністю кроків підвищення привілеїв.

До обмежень дослідження належить кілька аспектів. По-перше, перевага запропонованої моделі не є однаково великою для всіх типів сценаріїв: на сценаріях із плоским простором пошуку та високою часткою альтернативних OR-залежностей різниця між Proposed і StepPenalty помітно зменшується, а на окремих графах StepPenalty може демонструвати близькі або трохи кращі значення ТТС. По-друге, результати залежать від структури згенерованих MAL-графів та від вибору параметрів функції винагороди (λ, w_p, w_d) , хоча проведений аналіз чутливості показав наявність не вузької оптимальної точки, а відносно стійкої області параметрів.

Подальші дослідження доцільно спрямувати на два напрями. Перший пов'язаний з адаптивним налаштуванням параметрів λ , w_p і w_d безпосередньо в процесі навчання агента. Другий стосується розширення моделі на складніші багатосервісні вебсередовища з більш вираженою ієрархією цілей і багаторівневими залежностями між кроками атаки.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] M. C. Ghanem, and T. M. Chen, "Reinforcement Learning for Efficient Network Penetration Testing," *Information*, vol. 11, no. 1, p. 6, 2019. <https://doi.org/10.3390/info11010006>.
- [2] M. C. Ghanem, T. M. Chen, and E. G. Nepomuceno, "Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks," *Journal of Intelligent Information Systems*, vol. 60, no. 2, pp. 281-303, 2022. <https://doi.org/10.1007/s10844-022-00738-0>.
- [3] F. M. Zennaro, and L. Erdödi, "Modelling penetration testing with reinforcement learning using capture-the-flag challenges: Trade-offs between model-free learning and a priori knowledge," *IET Information Security*, vol. 17, no. 3, pp. 441-457, 2023. <https://doi.org/10.1049/ise2.12107>.
- [4] J. Yi, and X. Liu, "Deep Reinforcement Learning for Intelligent Penetration Testing Path Design," *Applied Sciences*, vol. 13, no. 16, p. 9467, 2023. <https://doi.org/10.3390/app13169467>.
- [5] А. Толкачова, та М.-М. Посувайло, «Тестування на проникнення з використанням глибокого навчання з підкріпленням», *Кибербезпека: освіта, наука, техніка*, т. 3, № 23, 2024. <https://doi.org/10.28925/2663-4023.2024.23.1730>.
- [6] В. Вікулов, та І. Пишнограв, «Автоматизоване виявлення вразливостей SQL-ін'єкцій у чат-ботах за допомогою навчання з підкріпленням», *Кибербезпека: освіта, наука, техніка*, т. 1, № 29, 2025. <https://doi.org/10.28925/2663-4023.2025.29.873>.
- [7] D. J. Leversage, and E. J. Byres, "Estimating a System's Mean Time-to-Compromise," *IEEE Security & Privacy*, vol. 6, no. 1, pp. 52-60, 2008. <https://doi.org/10.1109/msp.2008.9>.
- [8] W. Nzoukou, L. Wang, S. Jajodia, and A. Singhal, "A Unified Framework for Measuring a Network's Mean Time-to-Compromise," in *2013 IEEE 32nd International Symposium on Reliable Distributed Systems*, 2013, pp. 215-224. <https://doi.org/10.1109/srds.2013.30>.
- [9] U. Garg, G. Sikka, and L. K. Awasthi, "Empirical risk assessment of attack graphs using time to compromise framework," *International Journal of Information and Computer Security*, vol. 16, no. 1/2, p. 33, 2021. <https://doi.org/10.1504/ijics.2021.117393>.
- [10] E. Rencelj Ling, and M. Ekstedt, "Estimating Time-To-Compromise for Industrial Control System Attack Techniques Through Vulnerability Data," *SN Computer Science*, vol. 4, no. 3, 2023. <https://doi.org/10.1007/s42979-023-01750-z>.
- [11] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, *MITRE ATT&CK: Design and philosophy* (Technical Report MP180360R1). The MITRE Corporation, 2020. [Online]. Available: <https://www.mitre.org/sites/default/files/2021-11/prs-19-01075-28-mitre-attack-design-and-philosophy.pdf>.

- [12] P. Mell, K. Scarfone, and S. Romanosky, "Common Vulnerability Scoring System," *IEEE Security and Privacy Magazine*, vol. 4, no. 6, pp. 85-89, 2006. <https://doi.org/10.1109/msp.2006.145>.
- [13] M. Ibrahim, and R. Elhafiz, "Security Analysis of Cyber-Physical Systems Using Reinforcement Learning," *Sensors*, vol. 23, no. 3, p. 1634, 2023. <https://doi.org/10.3390/s23031634>.
- [14] M. Ibrahim, and R. Elhafiz, "Security Assessment of Industrial Control System Applying Reinforcement Learning," *Processes*, vol. 12, no. 4, p. 801, 2024. <https://doi.org/10.3390/pr12040801>.
- [15] Y. Wang, Y. Li, X. Xiong, J. Zhang, Q. Yao, and C. Shen, "DQfD-AIPT: An Intelligent Penetration Testing Framework Incorporating Expert Demonstration Data," *Security and Communication Networks*, vol. 2023, pp. 1-15, 2023. <https://doi.org/10.1155/2023/5834434>.
- [16] Л. Д. Ганенко, «Метод адаптивного формування винагороди за умов невизначеності динамічних об'єктів», *Телекомунікаційні та інформаційні технології*, № 1, с. 23-30, 2026. <https://doi.org/10.31673/2412-4338.2026.019003>.
- [17] B.-S. Kim, H.-W. Suk, Y.-H. Choi, D.-S. Moon, and M.-S. Kim, "Optimal Cyber Attack Strategy Using Reinforcement Learning Based on Common Vulnerability Scoring System," *Computer Modeling in Engineering & Sciences*, vol. 141, no. 2, pp. 1551-1574, 2024. <https://doi.org/10.32604/cmescs.2024.052375>.
- [18] А. Притула, та Л. Куперштейн, «Моделювання сценаріїв кібератак як марковського процесу із семантично обмеженим простором дій», *Кібербезпека: освіта, наука, техніка*, № 33, с. 555-569, 2026. <https://doi.org/10.28925/2663-4023.2026.33.1232>.
- [19] P. Johnson, R. Lagerström, and M. Ekstedt, "A Meta Language for Threat Modeling and Attack Simulations," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pp. 1-8, 2013. <https://doi.org/10.1145/3230833.3232799>.
- [20] V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-537, 2015. <https://doi.org/10.1038/nature14236>.
- [21] *EnterpriseLang: Enterprise Language for the Meta Attack Language framework*. [Online]. Available: <https://github.com/mal-lang/enterpriseLang>.
- [22] *WebGoat: A deliberately insecure web application* [Online]. Available: <https://github.com/WebGoat/WebGoat>.

Рекомендована кафедрою захисту інформації ВНТУ

Дата надходження 24.04.2026

Дата прийняття після доопрацювання 18.08.2026

Дата прийняття до друку після рецензування 25.08.2026

Дата публікації 2.10.2026

Притула Андрій Вікторович — аспірант кафедри захисту інформації, e-mail: andrik.pritula@gmail.com, <https://orcid.org/0009-0006-9632-0712>;

Куперштейн Леонід Михайлович — канд. техн. наук, доцент кафедри захисту інформації, e-mail: kupershtein.lm@gmail.com, <https://orcid.org/0000-0001-6737-7134>.

Вінницький національний технічний університет, Вінниця

A. V. Prytula¹

L. M. Kupershtein¹

Reward Model in Reinforcement Learning for Web Application Security Testing

¹Vinnitsia National Technical University

The paper addresses the problem of reward function design for reinforcement learning in cyberattack scenario analysis, where traditional approaches within Markov decision processes, focused primarily on reaching a terminal state, insufficiently reflect the temporal dynamics of compromise and the varying importance of intermediate objectives. A reward model is proposed that combines a temporal component time-to-compromise (TTC) with semantic weights of intermediate goals, particularly actions related to privilege escalation and access to critical resources. Unlike conventional reward schemes, this approach produces a locally informative reinforcement signal that guides learning not only toward achieving compromise but also toward constructing time-efficient and structurally meaningful attack trajectories. The experimental study was conducted on 12 stochastically generated MAL-scenarios (7200 runs per method) based on open attack models. The results showed that the proposed model yields a substantially lower mean TTC compared to three baseline approaches (improvement ranging from 5 % to 21 %) with high statistical significance, while simultaneously increasing the rate of successful target state achievement and reducing the number of inadmissible actions. The resulting policies are also more interpretable: the share of targeted actions in the constructed trajectories increases, while auxiliary steps are displaced in favor of strategically significant privilege escalation actions. Sensitivity analysis confirmed the robustness of the results to the choice of model parameters. The practical significance of the work lies in the potential application of the proposed approach for identifying, ranking, and prioritizing the most dangerous web application compromise scenarios with respect to their temporal and structural criticality.

Keywords: markov decision process; cybersecurity; threat; attack graph; reinforcement learning; neural network; reward function; machine learning; artificial intelligence; modeling.

Prytula Andrii V. — Post-Graduate Student of the Chair of Information Protection, e-mail: andrik.pritula@gmail.com . <https://orcid.org/0009-0006-9632-0712>;

Kupershtein Leonid M. — Cand. Sc. (Eng.), Associate Professor of the Chair of Information Protection, e-mail: kupershtein.lm@gmail.com . <https://orcid.org/0000-0001-6737-7134>