МЕТОД ЗА ОЦЕНКА НА КИБЕРИНЦИДЕНТИ С ИЗКУСТВЕН ИНТЕЛЕКТ

ас. д-р. инж. **Петя Иванова Петрова**

Висше училище по телекомуникации и пощи, София

Резюме: *Разширяването на приложението на изкуствения интелект в информационни системи, облачни платформи, автоматизирани процеси и автономни агенти променя характеристиките на съвременните киберинциденти. Изкуственият интелект може да бъде едновременно обект на атака, средство за нейното осъществяване, автономен участник в последователност от действия или фактор за ускоряване и разпространение на инцидента. Това затруднява използването на традиционни подходи за оценяване, които не отчитат степента на автономност, достъпа до външни инструменти, загубата на човешки контрол и възможността компрометирането да се разпространява между взаимосвързани AI компоненти.*

В статията се предлага метод за класифициране и количествена оценка на киберинциденти, свързани с изкуствен интелект. Методът първоначално определя ролята на ИИ в инцидента, след което оценява неговата техническа тежест чрез набор от нормализирани показатели. Особено внимание се отделя на степента на автономност, потенциала за разпространение, въздействието върху данни и информационни ресурси, злоупотребата с права за достъп и степента на запазен човешки контрол. Въвежда се интегрална оценка, която комбинира базовата тежест на инцидента с корекции за съвместното влияние на автономността и разпространението и за загубата на човешки надзор. Приложимостта на метода се предвижда да бъде проверена чрез сценарии, представящи AI-assisted phishing, prompt injection, компрометиране на автономен AI агент и разпространение в multi-agent среда.

Ключови думи: *изкуствен интелект, киберинцидент, киберсигурност, автономен агент, prompt injection, оценка на инциденти, multi-agent systems.*

METHOD FOR ASSESSING CYBER INCIDENTS WITH ARTIFICIAL INTELLIGENCE

Asst. Prof. Eng. **Petya Ivanova Petrova**, Ph.D

University of Telecommunications and Post, Sofia

Abstract: *The expanding use of artificial intelligence in information systems, cloud platforms, automated processes, and autonomous agents changes the characteristics of contemporary cyber incidents. Artificial intelligence may simultaneously become a target of*

attack, a tool supporting malicious activity, an autonomous participant in a sequence of actions, or a factor accelerating and propagating an incident. This limits the applicability of traditional assessment approaches that do not explicitly account for autonomy, access to external tools, loss of human oversight, and propagation between interconnected AI components.

This paper proposes a method for classification and quantitative assessment of cyber incidents related to artificial intelligence. The method first determines the role of AI in the incident and then evaluates its technical severity through normalized indicators. Particular attention is given to the degree of autonomy, propagation potential, impact on data and information resources, privilege abuse, and the level of retained human oversight. An integrated score is introduced by combining the baseline incident severity with corrections reflecting the joint influence of autonomy and propagation and the loss of human control. The method is intended to be examined through scenarios involving AI-assisted phishing, prompt injection, compromise of an autonomous AI agent, and propagation in a multi-agent environment.

Keywords: *artificial intelligence, cyber incident, cybersecurity, autonomous agent, prompt injection, incident assessment, multi-agent systems.*

Въведение

Интегрирането на изкуствен интелект в корпоративни информационни системи, комуникационни платформи, облачни услуги и процеси за автоматизация променя начина, по който възникват, развиват се и се разпространяват киберинцидентите. Докато при традиционните информационни системи основният обект на защита са инфраструктурата, приложните компоненти, идентичностите и данните, при AI-базираните системи към тях се добавят моделите, контекстът, паметта, инструментите, механизмите за автономно вземане на решения и взаимодействието с други системи.

Развитието на автономните AI агенти допълнително променя рисковия профил. Съвременният агент може не само да генерира отговор, а да планира последователност от действия, да използва външни инструменти, да извиква API, да обработва файлове, да използва машинни идентичности и да взаимодейства с други агенти. OWASP Top 10 for Agentic Applications 2026 разглежда като съществени рискове hijacking на целите на агента, злоупотреба с инструменти, нарушения при идентичностите и привилегиите, непредвидено изпълнение и други рискове, произтичащи от автономността на агентните системи. [14]

Тези характеристики показват, че понятието **AI-свързан киберинцидент** не трябва да бъде ограничено само до инцидент, при който е атакуван модел с изкуствен интелект. Съществуват поне четири различни ситуации: ИИ може да бъде цел на атаката; може да бъде използван като средство за подпомагане на атакуващия; може да извършва автономни действия като част от инцидента; или да увеличи мащаба и скоростта на неговото разпространение.

Целта на настоящото изследване е да се разработи метод за класифициране и количествена оценка на киберинциденти, свързани с изкуствен интелект, който отчита както традиционното техническо въздействие, така и специфичните характеристики на AI участието в инцидента.

За постигане на поставената цел се определят следните задачи:

1. Да се определят основните начини, по които изкуственият интелект може да участва в киберинцидент.
2. Да се разработи класификация на AI-свързаните киберинциденти според ролята на ИИ.
3. Да се определят количествени показатели, характеризиращи техническата тежест, автономността, разпространението, въздействието върху данни и ресурси, злоупотребата с права и човешкия контрол.
4. Да се разработи математически механизъм за агрегиране на показателите в интегрална оценка.

Основната постановка на изследването е, че при равни други технически последствия киберинцидент с висока степен на автономност и висок потенциал за разпространение следва да получи по-висока оценка от инцидент, при който AI има само подпомагаща роля и действията остават под непосредствен човешки контрол.

1. Киберинциденти, свързани с изкуствен интелект

1. 1. Роля на ИИ в киберинцидента

За целите на метода AI-свързан киберинцидент се разглежда като събитие или последователност от събития, при които AI компонент участва по начин, който има значение за възникването, развитието, изпълнението или последствията от инцидента. Предлага се следната класификационна структура:

{I_T, I_E, I_A, I_P},
където:

(I_T) – изкуственият интелект е **цел** на атаката;

(I_E) – изкуственият интелект е **средство** за извършване или подпомагане на атаката;

(I_A) – изкуственият интелект е **автономен участник** в инцидента;
(I_P) – изкуственият интелект е **фактор за разпространение или усилване** на инцидента.

Категорията (I_T) включва инциденти, при които атаката е насочена непосредствено към AI система или нейни специфични компоненти. Такива могат да бъдат prompt injection, model manipulation, poisoning на входни данни или контекст, злоупотреба с паметта на агента, компрометиране на AI supply chain и други действия, насочени към промяна на поведението или достъпа до AI системата. MITRE ATLAS и OWASP LLM/Agentic guidance предоставят съвременни таксономии на подобни техники и рискове. [3, 9, 13, 15, 16]

Категорията (I_E) включва инциденти, при които AI не е непосредствено компрометиран, а се използва от атакуващия като инструмент. Това може да бъде автоматизирано генериране на фишинг съдържание, анализ на потенциални цели, ускорено създаване на зловредни скриптове, персонализиране на социално инженерство или подпомагане на други етапи от атаката. [4, 14]

При (I_A) AI системата участва самостоятелно в извършването на действия. Това е специфично за agentic AI, при който моделът разполага с инструменти, права и механизми за изпълнение. Ако злонамерена инструкция промени неговата цел или контекст, последващите действия могат да бъдат извършени без непосредствено човешко потвърждение. Именно тук автономността се превръща от функционална характеристика в фактор за киберриска. [14]

Категорията (I_P) се отнася до ситуации, в които AI увеличава скоростта, обхвата или способността на инцидента да се пренася между системи. Това е особено важно при multi-agent архитектури, при които един агент може да изпраща резултати към друг, да споделя обща памет или да използва общи ресурси. Компрометиран контекст в подобна среда може да създаде последователност: [14]

[A_1\ → A_2\ → API\ → D\ → A_3],
където (A_1,A_2,A_3) са AI агенти, (API) е външен програмен интерфейс, а (D) е информационен ресурс.

Класификациите не са взаимно изключващи се. Един инцидент може първоначално да бъде (I_T), ако AI агент е манипулиран чрез prompt injection, след което да премине към (I_A), ако агентът самостоятелно изпълни действия, и едновременно да прояви характеристики на (I_P), ако компрометирането се разпространи към други системи.

Следователно ролята на ИИ може да бъде представена не само чрез една категория, а чрез бинарен вектор:

[t_T,t_E,t_A,t_P],
където:

[t_j\in{0,1}]

Това позволява един инцидент да бъде определен например като:

[\mathbf{T}_{AI}=[1,0,1,1]]
което означава, че AI е бил едновременно цел, автономен участник и фактор за разпространение.

1. 2. Характеристики за количествена оценка

След определяне на ролята на ИИ инцидентът се описва чрез оценъчен вектор:

[S, A, P, D, V, R, H],
където:

- (S) – техническа тежест на инцидента;
- (A) – степен на автономност на AI компонента;
- (P) – потенциал за разпространение;
- (D) – въздействие върху данни и информационни ресурси;
- (V) – злоупотреба с права, идентичности и достъп;
- (R) – сложност на възстановяването;
- (H) – степен на загуба или заобикаляне на човешкия контрол.

Тези характеристики се избират така, че да не се дублират напълно традиционните показатели за severity, а да добавят свойства, специфични за AI-базираните системи.

Особено значение има (A). Стойността му трябва да бъде ниска, когато AI само подпомага човек, и висока, когато системата самостоятелно избира и изпълнява

действия. Аналогично (P) е нисък при локализиран инцидент и висок, когато AI има достъп до множество ресурси, агенти или автоматизирани workflow процеси. [14]

Показателят (H) представя човешкия надзор. При наличие на задължително human approval преди критично действие стойността е ниска. При възможност агентът да извършва действия без потвърждение, или когато механизмът за човешки контрол е заобиколен, стойността се увеличава. [7, 8, 14]

Комбинацията от (A), (P) и (H) е централна за предложения метод, защото именно тези параметри отличават автономния AI-свързан инцидент от традиционна атака, при която всички действия се управляват непосредствено от атакуващия.

2. Метод за оценка на киберинциденти, свързани с изкуствен интелект

Предложеният метод е предназначен за структурирана количествена оценка на киберинциденти, при които изкуственият интелект участва като обект на атаката, средство за нейното осъществяване, автономен участник или фактор за разпространение. Методът използва вече определените характеристики на инцидента и ги обединява в последователна процедура, която позволява едновременно да бъдат отчетени техническите последствия и специфичното влияние на AI компонента.

Основната идея е оценяването да се реализира на две равнища. Първото определя **базовата техническа тежест на инцидента**, независимо от ролята на ИИ. Второто отчита допълнителното влияние, което произтича от автономното поведение на AI системата, способността ѝ да разпространява инцидента и степента, в която човешкият контрол е ограничен или заобиколен.

По този начин се избягва автоматичното приемане, че всеки инцидент с участие на ИИ е по-критичен от традиционен киберинцидент. Повишаване на крайната оценка се допуска само когато AI-специфичните характеристики действително увеличават въздействието или усложняват контрола върху инцидента.

Методът се реализира чрез последователността:

M1→M2→M3→M4→M5→M6,

където:

- M1 – определяне на ролята на ИИ в инцидента;
- M2 – количествено оценяване на техническите последствия;
- M3 – количествено оценяване на AI-специфичните характеристики;
- M4 – формиране на базова и AI-коригирана оценка;
- M5 – определяне на крайната стойност и класифициране на инцидента;
- M6 – определяне на приоритет за реакция.

2. 1. Определяне на ролята на ИИ

В първата стъпка се използва вече дефинираният бинарен вектор:

$TAI=[tT,tE,tA,tP]$,

където:

$tj \in \{0,1\}$.

Той определя дали изкуственият интелект е:

- цел на атаката;
- средство за нейното осъществяване;
- автономен участник;
- фактор за разпространение.

Класификационният вектор не участва директно в интегралната формула. Неговата функция е да характеризира **механизма на участие на ИИ**, докато количествената тежест се определя чрез параметрите на оценъчния вектор.

Това разграничение е методологично важно. Например два инцидента могат да имат приблизително еднаква крайна стойност, но при единия ИИ да е използван само като средство за създаване на фишинг съдържание, а при другия компрометиран автономен агент да извършва действия в информационната инфраструктура. Крайната числова стойност в тези случаи трябва да бъде допълвана от класификационния профил.

2. 2. Нормализиране на показателите

За да бъдат различните характеристики съпоставими, всички показатели се нормализират в интервала:

$xi \in [0,1]$.

По-високата стойност винаги означава по-неблагоприятно състояние.

За практическото оценяване се използва петстепенна скала:

Стойност	Интерпретация
0	липсващо или пренебрежимо въздействие
0.25	Ниско
0.5	Средно
0.75	Високо
1	Критично

След агрегиране отделните компонентни оценки могат да приемат и междинни стойности. Тази нормализация позволява параметри с различна техническа природа да бъдат включени в единен изчислителен механизъм.

2. 3. Базова техническа оценка

От вече определения вектор

[S,A,P,D,V,R,H]

четири показателя описват основно техническите последствия:

B=[S,D,V,R].

Показателите S, D, V и R представят съответно техническата тежест на инцидента, въздействието върху данни и информационни ресурси, злоупотребата с идентичности и права и сложността на възстановяването. При AI-базирани системи оценяването следва да обхваща не само традиционните активи, но и системни инструкции, контекст, памет, входни и изходни данни, машинни идентичности, API достъп и резултати от модела. Възстановяването може да изисква ротация на ключове и токени, проверка на изпълнените действия, изчистване на компрометиран контекст и повторна валидация на генерираните резултати [6, 13, 14].

2. 4. Формиране на базовата стойност

Базовата техническа оценка се определя чрез претеглено агрегиране:

CB=wSS+wDD+wVV+wRR,

при:

wS+wD+wV+wR=1,

и:

wi≥0.

Следователно:

CB∈[0,1].

При липса на достатъчно емпирични основания за различна значимост на показателите в експерименталната проверка могат да се използват равни тегла:

wS=wD=wV=wR=0.25.

Тогава:

CB=(S+D+V+R)/4.

Получената стойност характеризира техническите последствия от инцидента **преди да бъде отчетено специфичното влияние на изкуствения интелект**.

Това е важно, защото позволява да бъде направено сравнение:

Традиционна техническа тежест срещу AI-коригиранатежест.

2. 5. AI-специфична оценка

Останалите три показателя от оценъчния вектор:

[A,P,H]

характеризират специфичните свойства на AI участието.

Те не описват директно вече настъпилите технически последствия, а способността на системата да продължи, разшири или автономно развие инцидента.

Автономност A

Параметърът:

$A \in [0,1]$

определя степента, в която AI компонентът може самостоятелно да избира и изпълнява действия.

A	Характеристика
0.00	AI няма самостоятелна изпълнителна функция
0.25	AI генерира препоръки или съдържание
0.50	изпълнява предварително ограничени действия
0.75	планира и изпълнява последователни действия
1.00	извършва критични действия без предварително човешко одобрение

Така например използване на ИИ за генериране на phishing текст не следва да получава същата стойност като автономен агент, който може самостоятелно да извиква API и да изменя информационни ресурси.

Потенциал за разпространение P

Параметърът:

$P \in [0,1]$

представя възможността инцидентът да се разпространи отвъд първоначално засегнатия компонент.

P	Характеристика
0.00	инцидентът е локализиран
0.25	ограничено въздействие върху един свързан ресурс
0.50	възможно въздействие върху няколко свързани ресурса
0.75	възможно разпространение между множество системи или агенти
1.00	висок потенциал за автоматизирано системно разпространение

Параметърът не представя вече настъпилото въздействие, а **способността на инцидента да разшири своя обхват**.

Загуба на човешки контрол H

Показателят:

$H \in [0,1]$

характеризира степента, в която механизмите за човешки надзор са отслабени или заобиколени.

H	Характеристика
0.00	всички значими действия се одобряват от човек
0.25	автоматизирани са само нискорискови операции
0.50	част от значимите действия се извършват без предварително одобрение
0.75	критични действия могат да бъдат автоматизирани
1.00	липсва или е заобиколен ефективният човешки контрол

Важно е H да не бъде отъждествяван с A. Една система може да бъде силно автономна, но да използва надежден механизъм за human approval преди критични операции.

2. 6. Коефициент на AI участие

Самостоятелното използване на:

AP

не е достатъчно, защото може да подцени случай, при който автономността е много висока, но разпространението е ограничено, или обратно.

Затова се въвежда:

$QAI=\alpha A+\gamma P+\delta AP,$

при:

$\alpha+\gamma+\delta=1,$

и:

$\alpha,\gamma,\delta\geq 0.$

Следователно:

$QAI\in[0,1].$

Членовете:

αA

и

γP

представят самостоятелното влияние на автономността и разпространението, докато:

δAP

представя тяхното взаимодействие.

За експерименталната проверка може да бъде използвано:

$\alpha=0.30,\gamma=0.30,\delta=0.40.$

По-високата стойност на δ позволява да се отчете по-силно ситуацията, при която AI компонентът едновременно е силно автономен и има възможност да разшири инцидента.

Тези стойности са **експериментални параметри**, а не универсално определени константи.

2. 7. AI-коригирана стойност

След определяне на базовата техническа оценка CB и коефициента QAI се формира:

$$CA=CB+\lambda QAI(1-CB),$$

където:

$$0\leq\lambda\leq1.$$

Параметърът λ контролира максималното влияние на AI-специфичните характеристики.

Формулата удовлетворява:

$$CB\leq CA\leq1.$$

Когато:

$$QAI=0,$$

получаваме:

$$CA=CB.$$

Следователно AI участието не увеличава автоматично оценката.

Когато:

$$QAI\rightarrow1,$$

AI-специфичната корекция нараства, но крайният резултат остава ограничен в интервала [0,1].

2. 8. Отчитане на човешкия контрол

Последният AI-специфичен параметър е H.

Крайната стойност се определя чрез:

$$CF=CA+\mu H(1-CA),$$

където:

$$0\leq\mu\leq1.$$

Следователно:

$$CA\leq CF\leq1.$$

Общият математически израз на метода е:

$$CF=CB+\lambda(\alpha A+\gamma P+\delta AP)(1-CB)+\mu H[1-CB-\lambda(\alpha A+\gamma P+\delta AP)(1-CB)]$$

Това уравнение представлява основното **изчислително ядро на предложения метод**.

В него:

CB

отразява традиционната техническа тежест;

QAI

представя специфичното влияние на автономността и разпространението;

a:

H

отчита загубата на човешки контрол.

2. 9. Класифициране на крайния резултат

Крайната стойност:

$CF \in [0,1]$

се интерпретира чрез пет нива:

CF	Ниво на инцидента
$0.00 \leq CF < 0.20$	много ниско
$0.20 \leq CF < 0.40$	ниско
$0.40 \leq CF < 0.60$	средно
$0.60 \leq CF < 0.80$	високо
$0.80 \leq CF \leq 1.00$	критично

Резултатът трябва да се представя заедно с класификационния вектор:

$IAI* = (TAI, CF)$.

Например:

$IAI* = ([1,0,1,1], 0.86)$

означава **критичен AI-свързан киберинцидент**, при който системата с изкуствен интелект е била обект на атаката, участвала е автономно и е допринесла за разпространението.

3. Експериментална проверка на предложения метод

3. 1. Организация на експерименталната проверка

За проверка на поведението на предложения метод се използват четири сценария, които представят различна роля и различна степен на участие на изкуствения интелект в развитието на киберинцидента. Целта не е да се симулира конкретна реална инфраструктура, а да се установи дали методът формира различими крайни оценки при

промяна на техническите последствия, автономността, потенциала за разпространение и степента на човешки контрол.

Сценариите са подбрани така, че да представят постепенно усложняване:

IE→IT→IT+IA→IT+IA+IP.

Използват се равни тегла за базовите технически показатели:

wS=wD=wV=wR=0.25.

За AI-специфичния коефициент се приемат:

$\alpha=0.30, \gamma=0.30, \delta=0.40$.

За влиянието на AI-специфичната корекция и загубата на човешки контрол се използват:

$\lambda=0.35, \mu=0.20$.

Стойностите са експериментални и са избрани за проверка на поведението на метода. Те не представляват универсално валидирани константи.

3. 2. Експериментални сценарии

Сценарий 1 – AI-assisted phishing

Изкуственият интелект се използва за създаване и персонализиране на фишинг съдържание. AI компонентът не изпълнява самостоятелни действия в целевата информационна система и няма непосредствен достъп до допълнителни ресурси. Ролята му е преимуществено подпомагаща:

TAI(1)=[0,1,0,0].

Сценарий 2 – Prompt injection срещу корпоративен AI assistant

Злонамерена инструкция променя поведението на AI assistant и води до неправомерно извеждане на информация. Инцидентът е насочен към AI системата, но автономното изпълнение остава ограничено:

TAI(2)=[1,0,0,0].

Сценарий 3 – Компрометиран автономен AI агент

Манипулиран контекст променя целта на агент, който има достъп до външни инструменти и машинни идентичности. Агентът извършва последователни действия без непосредствено човешко потвърждение:

TAI(3)=[1,0,1,0].

Сценарий 4 – Разпространение в multi-agent среда

Компрометиран AI агент предава манипулиран контекст към други агенти и използва свързани API и информационни ресурси. Налице са висока автономност,

висок потенциал за разпространение и практически липсващ непосредствен човешки контрол:

TAI(4)=[1,0,1,1]. Входните стойности са систематизирани в Таблица 1.

Таблица 1

Параметри на експерименталните сценарии

Показател	Сценарий 1	Сценарий 2	Сценарий 3	Сценарий 4
S	0.25	0.5	0.75	0.75
D	0.25	0.5	0.5	0.75
V	0.25	0.25	0.75	1
R	0.25	0.5	0.75	0.75
A	0.25	0.25	0.75	1
P	0.25	0.25	0.5	1
H	0.25	0.25	0.75	1

За илюстриране на възможностите по сценариите ще разгледаме накратко Сценарий 3.

3. 3. Изчисление за Сценарий 3

Последователността на метода подробно се изчислява в Сценарий 3. Базовата техническа оценка е:

CB=(0.75+0.50+0.75+0.75)/4=0.6875.

AI-специфичният коефициент е:

QAI=0.30(0.75)+0.30(0.50)+0.40(0.75)(0.50).

Следователно:

QAI=0.225+0.150+0.150=0.525.

AI-коригираната оценка е:

CA=0.6875+0.35(0.525)(1−0.6875).

Получава се:

CA≈0.745.

При:

H=0.75

крайната оценка е:

CF=0.745+0.20(0.75)(1−0.745).

Следователно:

CF≈0.783.

Получената стойност се намира непосредствено под границата за критичен инцидент и съответства на **високо ниво**.

Тъй като:

A=0.75

и:

H=0.75,

инцидентът изисква повишено внимание, въпреки че P=0.50 не достига прага за висока AI-специфична критичност.

4. Резултати от четирите сценария

Резултатите са представени в Таблица 2.

Таблица 2

Резултати от експерименталната проверка

Сценарий	CB	QAI	CA	CF	Ниво
1	0.25	0.175	0.296	0.331	ниско
2	0.438	0.175	0.472	0.498	средно
3	0.688	0.525	0.745	0.783	високо
4	0.813	1	0.878	0.903	критично

Резултатите показват последователно нарастване на крайната оценка: от 0.331 при AI-assisted phishing и 0.498 при prompt injection до 0.783 при компрометиран автономен агент и 0.903 при multi-agent разпространение. Така методът различава както техническата тежест, така и допълнителното влияние на автономността, разпространението и човешкия контрол.

3. 5. Влияние на AI-специфичната корекция

За да се оцени реалното влияние на AI характеристиките, може да се сравнят CB и CF.

При Сценарий 1 разликата е:

0.331−0.250=0.081.

При Сценарий 2:

0.498−0.438=0.060.

При Сценарий 3:

0.783−0.688=0.095.

При Сценарий 4:

0.903−0.813=0.090.

Резултатите показват, че AI-корекцията не действа като постоянна добавка. Нейната стойност зависи от комбинацията между базовата техническа тежест, автономността, разпространението и човешкия контрол.

Това свойство е важно, защото методът не приема автоматично, че участието на ИИ увеличава тежестта с фиксирана стойност.

3. 6. Проверка на влиянието на автономността и разпространението

За допълнителна проверка се разглежда Сценарий 3 при запазване на техническите показатели, но при намаляване на автономността:

A:0.75→0.25.

При непроменени:

P=0.50, H=0.75

AI-специфичният коефициент намалява.

Това води до по-ниска крайна оценка, без да се променя:

CB=0.6875.

Следователно изменението на CF може да бъде отнесено единствено към промяната в AI-специфичните характеристики.

Същата логика може да бъде приложена към потенциала за разпространение. Ако при Сценарий 4:

P:1.00→0.25,

при запазване на останалите технически последствия, QAI намалява значително и крайната оценка се понижава.

Това потвърждава основната постановка на метода:

ΔA>0 или ΔP>0

при равни други условия води до:

ΔCF>0.

3. 7. Извод от експерименталната проверка

Четирите сценария показват, че предложеният метод формира различими оценки при различни роли на изкуствения интелект и при различни стойности на автономност, разпространение и човешки контрол.

Получената последователност е:

0.331<0.498<0.783<0.903.

Тя съответства на преминаването от подпомагащо участие на ИИ към компрометиране на AI система, автономно изпълнение и накрая – разпространение в multi-agent среда.

Експерименталната проверка показва също, че AI-специфичната корекция не замества техническата оценка, а я допълва. Така два инцидента с близки технически последствия могат да бъдат разграничени според степента на автономност и възможността инцидентът да продължи или да се разпространи без непосредствен човешки контрол.

4. Анализ на резултатите, приложимост и ограничения

Получените резултати показват, че разграничаването между базовата техническа оценка CB и крайната стойност CF позволява да се отделят непосредствените технически последствия от специфичното влияние на ИИ. Автономността е значима, когато AI компонентът може да планира и изпълнява действия без непосредствено човешко потвърждение; потенциалът за разпространение е съществен при достъп до други агенти, API, споделена памет или автоматизирани workflow процеси; а H отчита доколко тези действия остават под ефективен human-in-the-loop контрол [7, 8, 14].

Практически методът може да се използва при incident triage и SOC анализ, когато входните стойности се извличат от SIEM/XDR събития, IAM/PAM системи, audit trails, записи за извикани инструменти и API, machine identities и agent telemetry. Класификационният вектор TAI допълва числовата стойност CF и позволява близки по тежест инциденти да бъдат разграничени според механизма на AI участие [11, 12, 14].

Ограничения на метода

Предложеният метод има няколко ограничения, които трябва да бъдат отчетени при интерпретирането на резултатите.

На първо място, стойностите на показателите зависят от качеството на наличната информация за инцидента. При непълни логове, липсваща telemetry или недостатъчна проследимост на действията на AI агента част от параметрите могат да бъдат оценени с по-ниска точност.

На второ място, теглата:

$w_i, \alpha, \gamma, \delta$

и параметрите:

λ, μ

не са универсални. В настоящото изследване те се използват за експериментална проверка и следва да бъдат калибрирани върху реални набори от инциденти. [1, 7, 8]

На трето място, разграничението между неправилен AI резултат и киберинцидент може да бъде сложно. Не всяка халюцинация, неточност или непредвидено поведение представлява инцидент със сигурността. Необходима е връзка

с нарушение на контрол, неправомерно действие, компрометиране или въздействие върху защитени ресурси. [13, 14, 16]

На четвърто място, multi-agent системите могат да бъдат значително по-сложни от разглежданата линейна схема на разпространение. При реални архитектури е възможно да има циклични взаимодействия, споделена памет и динамично създавани агенти. [14]

Поради това методът следва да бъде разглеждан като структуриран количествен подход за първична оценка и сравнение, а не като заместител на техническия forensic анализ или пълния incident response процес.

Изводи

Проведеното изследване позволява да бъдат формулирани следните основни изводи:

1. Киберинцидентите, свързани с изкуствен интелект, не могат да бъдат оценявани само чрез традиционното техническо въздействие. При системите с изкуствен интелект допълнително значение имат степента на автономност, потенциалът за разпространение и степента на запазен човешки контрол.

2. Ролята на ИИ в инцидента следва да бъде разграничена от неговата тежест. Предложеният класификационен вектор

TAI=[tT,tE,tA,tP]

позволява да бъде определено дали ИИ е цел на атаката, средство за нейното осъществяване, автономен участник или фактор за разпространение. Това е важно, тъй като един и същ инцидент може да включва повече от една роля на ИИ.

3. Разделянето на оценката на базова и AI-специфична част предотвратява необоснованото завишаване на тежестта само поради наличие на изкуствен интелект. Базовата стойност CB представя техническите последствия, докато AI-корекцията отчита специфичните характеристики на автономното поведение.

4. Комбинацията между автономност и разпространение има по-съществено значение от самостоятелното им разглеждане. При наличие на висока автономност и възможност за взаимодействие с множество системи, API или други агенти инцидентът може да се развива без последователно управление от страна на атакуващия. Актуалните OWASP материали за agentic AI разглеждат именно автономните действия, tool misuse, identity and privilege abuse и взаимодействието между агентни компоненти като съществени области на сигурността. [14]

5. Човешкият контрол представлява самостоятелен фактор при оценяването. Високата автономност не води задължително до критичен инцидент, когато критичните действия са ограничени чрез ефективен human-in-the-loop механизъм. Обратно, заобикалянето на този контрол увеличава способността на AI компонента да продължи нежелано поведение.

Заключение

Навлизането на генеративни модели, AI assistants и автономни AI агенти в информационните системи формира нова среда за възникване и развитие на киберинциденти. Съвременните AI компоненти могат да използват инструменти, програмни интерфейси, машинни идентичности и външни информационни ресурси, а при agentic архитектури могат да планират и изпълняват последователност от действия. Тази промяна налага разширяване на традиционните подходи за оценяване на киберинциденти. [5, 8, 14]

В статията е предложен метод за класифициране и количествена оценка на киберинциденти, свързани с изкуствен интелект. Методът разглежда ролята на ИИ чрез четири основни проявления – цел на атаката, средство за нейното реализиране, автономен участник и фактор за разпространение. Подобно разграничение позволява да се представи по-точно реалното участие на AI компонента и да се отчетат случаи, при които един инцидент едновременно притежава няколко от тези характеристики.

Количествената част на метода разграничава базова техническа тежест и AI-специфично влияние. Базовата оценка отчита въздействието върху системи, данни, идентичности и възстановяване, а корекцията включва автономността, потенциала за разпространение и загубата на човешки контрол. Сценарийната проверка формира последователно нарастващи стойности и показва чувствителност на метода към тези характеристики [14].

Научният принос на изследването се изразява в предлагането на структуриран метод, който отделя техническите последствия от специфичните фактори, произтичащи от автономното участие на ИИ, и ги обединява в интегрална количествена оценка. Предложената класификация и математическа зависимост създават възможност за сравнение на различни категории AI-свързани инциденти при единна нормализирана скала.

Практическото приложение на метода може да бъде насочено към процесите по incident triage, SIEM/XDR анализ, SOC операции, управление на AI agents и определяне на приоритет при реагиране. В бъдещо развитие е необходимо валидиране върху реални набори от инциденти, емпирично калибриране на тегловните коефициенти и разработване на автоматизирани механизми за извличане на показателите от agent telemetry, audit logs и системи за управление на идентичности. [11, 12, 14]

Литература

[1] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1.
[2] Chang, H., Bao, E., Luo, X., & Yu, T. (2026). Overcoming the Retrieval Barrier: Indirect Prompt Injection in the Wild for LLM Systems. 35th USENIX Security Symposium (USENIX Security 26).
[3] Chen, S., Piet, J., Sitawarin, C., & Wagner, D. (2025). StruQ: Defending Against Prompt Injection with Structured Queries. 34th USENIX Security Symposium (USENIX Security 25), 2383–2400.
[4] ENISA. (2025). ENISA Threat Landscape 2025. European Union Agency for Cybersecurity.
[5] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
[6] ISO/IEC. (2022). ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements. International Organization for Standardization.
[7] ISO/IEC. (2023). ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. International Organization for Standardization.
[8] ISO/IEC. (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. International Organization for Standardization.
[9] Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). Formalizing and Benchmarking Prompt Injection Attacks and Defenses. 33rd USENIX Security Symposium (USENIX Security 24).
[10] Liu, F. W., & Hu, C. (2024). Exploring Vulnerabilities and Protections in Large Language Models: A Survey. arXiv preprint arXiv:2406.00240.
[11] MITRE. (2026). Adversarial Threat Landscape for Artificial-Intelligence Systems – ATLAS. MITRE Corporation.
[12] NIST. (2024). Cybersecurity Framework (CSF) 2.0. National Institute of Standards and Technology.
[13] OWASP Foundation. (2026). OWASP GenAI LLM Top 10 2026. OWASP GenAI Security Project.
[14] OWASP Foundation. (2026). OWASP Top 10 for Agentic Applications 2026. OWASP GenAI Security Project.
[15] Rababah, B., Shang, Wu, Kwiatkowski, M., Leung, C., & Akcora, C. G. (2024). SoK: Prompt Hacking of Large Language Models. arXiv preprint arXiv:2410.13901.
[6] Rehberger, J. (2024). Trust No AI: Prompt Injection Along the CIA Security Triad. arXiv preprint arXiv:2412.06090.
[17] Shao, Z., Fleming, C., & Baluta, T. (2026). Cordyceps: Covert Control Attacks on LLMs via Data Poisoning. 35th USENIX Security Symposium (USENIX Security 26).
[18] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1.
[19] Zhang, M., Jia, Y., Tan, Z., Jiang, S., Gong, N. Z., Chen, T., & Song, D. (2026). Measuring Real-World Prompt Injection Attacks in LLM-based Resume Screening. 35th USENIX Security Symposium (USENIX Security 26).