Studia i Materiały Informatyki Stosowanej, Tom 16, Nr 3, 2024
str. 24-33

AI system in the context of: threat modeling, its risk management and regulatory requirements

Emanuel Krzysztoń

Kazimierz Wielki University, Faculty of Computer Science
Kopernika 1, 85-074 Bydgoszcz
emanuel.krzyszton@ukw.edu.pl

Abstract:*The work presents a comprehensive approach to threat modelling and risk management in AI (Artificial Intelligence) systems. Providing methods, tools and guidance for organisations in building resilient, legally compliant AI systems. The proposed systematic approach enables the identification and minimisation of the consequences of threats in AI systems, opening up new research directions in the field of artificial intelligence security.*

Słowa kluczowe:*Artificial intelligence; Machine learning; Threat modeling, risk management, AI Act.*

System AI w kontekście: modelowania zagrożeń, zarządzania jego ryzykiem i wymagań regulacyjnych

Streszczenie: *Praca przedstawia kompleksowe podejście do modelowania zagrożeń i zarządzania ryzykiem w systemach SI (Sztuczna Inteligencja). Dostarczając metody, narzędzia i wskazówki dla organizacji w budowaniu odpornych na zagrożenia systemów SI, zgodnych z prawem. Proponowane systemowe podejście umożliwia identyfikację i minimalizację następstw zagrożeń w systemach SI, otwierając nowe kierunki badań w dziedzinie bezpieczeństwa sztucznej inteligencji.*

Słowa kluczowe*: Sztuczna inteligencja; Uczenie maszynowe; modelowanie zagrożeń, zarządzanie ryzykiem, AI Act.*

1. Wprowadzenie

Rozwój sztucznej inteligencji (SI) niesie ze sobą ogromny potencjał, jednak wiąże się również z nowymi wyzwaniami w zakresie bezpieczeństwa. Najnowsze badania [1-3] potwierdzają, że zagrożenia, rozumiane jako potencjalny czynnik mogący spowodować przerwę w działaniu lub awarię systemu SI, są zróżnicowane i mogą wystąpić na każdym etapie jego cyklu życia [4]**.**Na tej podstawie możemy podzielić czynniki wpływające na działanie systemów SI ze względu na ich charakter.Wyróżniamy czynniki zależne od człowieka (świadome i przypadkowe) oraz niezależne (środowiskowe). To podejście pozwala dostrzec, że wraz ze wzrostem złożoności systemów SI narasta wiele problemów. Do zagrożeń związanych ze sztuczną inteligencją należą: nieprecyzyjne definiowanie celów, co może prowadzić do nieoczekiwanych konsekwencji i utraty kontroli nad systemem SI. Dodatkowo, złożoność modeli SI oraz stosowanie niewłaściwych metryk do oceny ich skuteczności stanowią poważne wyzwanie. Problemy związane z jakością danych treningowych, takie jak uprzedzenia i niewystarczająca różnorodność, mogą prowadzić do błędnych decyzji. Ponadto, naruszenia prywatności oraz cyberataki zagrażają bezpieczeństwu danych i kontroli nad systemem SI.W 2024 roku Unia Europejska przyjęła pierwsze kompleksowe przepisy regulujące SI (AI Act) [5]. Celem tej regulacji jest stworzenie wspólnych ram prawnych, które zapewnią bezpieczeństwo, przejrzystość oraz poszanowanie podstawowych praw i wolności obywateli w związku z rozwojem i wykorzystywaniem systemów SI. Jednak jednym z głównych wyzwań w implementacji tych przepisów jest niepewność związana z klasyfikacją ryzyka systemów AI. Kryteria kategoryzacji systemów na różne poziomy ryzyka są otwarte na interpretację, prowadząc do potencjalnych przeszkód dla organizacji. Badanie [6] przeprowadzone na 100 organizacjach potwierdza tę tezę, wykazując, że aż 40% z nich nie zostało jednoznacznie sklasyfikowanych ze względu na niejasności w regulacjach.

1.1. Cel badawczy i metodologia

Głównym celem badawczymjest wskazanie podejścia do modelowania zagrożeń spełniającego wymagania regulacyjne w zakresie zarządzania ryzykiem w systemie AI. W tym celu zostanie wykorzystana metoda projektowania systemów informatycznych poprzez tzw. podejście systemowe [7].

1.2. Struktura artykułu

W pracy przedstawiono kompleksowe podejście do modelowania zagrożeń i zarządzania ryzykiem w systemach SI, które pozwala na budowę rozwiązań bezpiecznych i zgodnych z obowiązującymi regulacjami, takimi jak AI Act [5]. Opierając się na międzynarodowych normach ISO [4,8-14], analizie cyklu życia systemu SI oraz wybranych metodach modelowania zagrożeń [15-24], zaproponowano systemowe podejście, które umożliwia identyfikację i minimalizację potencjalnych zagrożeń na wczesnym etapie rozwoju systemu SI. Ponadto, zwrócono uwagę na znaczenie współpracy między ekspertami z różnych dziedzin oraz na konieczność ciągłego dostosowywania strategii zarządzania ryzykiem do dynamicznie zmieniającego się środowiska technologicznego.Badanie przyczynia się do rozwoju kompleksowego modelu zapewnienia bezpieczeństwa systemu AI, umożliwiającego organizacjom tworzenie innowacyjnych rozwiązań przy zachowaniu najwyższych standardów bezpieczeństwa.

W sekcji 2 opisano wymagania regulacyjne dla systemów SI. W sekcji 3 przedstawiono narzędzia do zarządzania ryzykiem w odniesieniu do norm. W sekcji 4 opisano cykl życia systemu SI w odniesieniu do normy. W sekcji 5 omówiono wybrane metody modelowania zagrożeń oraz dokonano analizy potencjalnych zagrożeń w ramach wybranego cyklu życia systemu SI. W sekcji 6 dokonano analizy SWOT proponowanego podejścia systemowego oraz wskazano ograniczenia dla tego podejścia. Sekcja 7 zawiera wnioski oraz dalsze kierunki badań.

2. Wymagania regulacyjne dla systemów SI

Pierwsze sformalizowane definicje systemuSI pojawiły się już w normie [4]. Jednakże, wraz z dynamicznym rozwojem tejtechnologii, konieczne stało się stworzenie bardziej kompleksowych ram prawnych. Unia Europejska (UE), w odpowiedzi na te potrzeby, wprowadziłaAI Act [5].Ten akt prawny definiuje system SI jako „system maszynowy zaprojektowany do działania z różnym stopniem autonomii po wdrożeniu, zdolny do adaptacji i generowania na podstawie danych wejściowych wyników o charakterze predykcyjnym, treściowym, rekomendacyjnym lub decyzyjnym, które mogą oddziaływać na środowisko fizyczne lub wirtualne".W ramach AI Act systemy sztucznej inteligencji zostały poddane rygorystycznej klasyfikacji, obejmującej trzy kategorie (Tabela 1). Podejście proponuje zachowanie proporcjonalność do potencjalnych zagrożeń wynikających z zastosowania SI.

Tabela 1. *Regulacja - podział systemów SI wg. [5] (opracowanie własne).*

System SI	Zgodnie z [5]	Szczegółowo
Niedozwolony	Rozdział II artykuł 5	Zakazane praktyki związane z AI, m.in. dark patterns, micro-targeting, social scoring czy zdalną identyfikacją biometryczną w czasie rzeczywistym.
Wysokiego ryzyka	Rozdział III artykuł 6, 8-15	Wymagania dla systemów AI wysokiego ryzyka, obejmujące m.in. zarządzanie ryzykiem i monitoring, mając na celu zapewnienie bezpieczeństwa i ochrony praw użytkowników.
Ograniczonego ryzyka	Rozdział IV artykuł 50	Wykorzystanie systemów AI w interakcji z użytkownikiem powinno być zawsze przejrzyste, a użytkownik powinien mieć prawo do decydowania o kontynuowaniu lub przerwaniu.

Rozporządzenie szczegółowo określa zakazane praktyki związane z systemami SI, które nie mogą być wykorzystywane ani udostępniane na terenie Unii Europejskiej (art. 5). Chociaż rozróżnienie to może nie być oczywiste, kluczowe jest zrozumienie, że zakaz dotyczy sposobu, w jaki SI jest używana, a nie samej technologii[25]. Artykuł 5 precyzyjnie określa m.in. manipulację behawioralną (z ang. dark patterns), czyli

technikimanipulacyjne do wpływania na decyzję i zachowania użytkowników bez ich świadomej zgody [26], słabość personalizowanego odbiorcy (z ang. microtargeting) czyli próby oszukania określonych grup ludzi do osiągnięcia nieetycznych celów [27], klasyfikowanie i ocenianie ludzi na podstawie zachowań społecznych (z ang. social scoring) prowadzące do dyskryminacji lub ograniczenia wolności [28], nieograniczony monitoring w „czasie rzeczywistym" do identyfikacji biometrycznej użytkownika w miejscach publicznych, naruszający prywatność i wolności obywatelskie [29]. Wprowadzenie na rynek systemów SI wysokiego ryzyka (art. 6) jest możliwe jedynie po spełnieniu wymagań określonych w art. 8-15. Dla pozostałych systemów SI, które nie mieszczą się w powyższych kategoriach, art. 50 nakłada obowiązek informowania użytkowników o interakcji z systemem SI, umożliwiając im świadomą decyzję o kontynuowaniu lub przerwaniu takiej interakcji.

3. Zarządzanie ryzykiem w systemach SI

Artykuł 9 i 10 rozporządzenia AI Act [5] nakłada na twórców i użytkowników systemów SI sklasyfikowanych w grupie o wysokim ryzyku obowiązek zarządzania ryzykiem związanym z ich zastosowaniem. Spełnienie tego wymogu jest możliwe poprzez:

- implementację i utrzymanie systemów zarządzania ryzykiem w celu identyfikacji, oceny i ograniczenia zagrożeń,
- dokumentację i aktualizację procesów zarządzania ryzykiem,
- ciągłe doskonalenie systemów SI poprzez testowanie, walidację i udoskonalanie zbiorów danych.

Zgodnie z powyższym zarządzanie ryzykiem jest kluczowym procesem identyfikacji, oceny i przeciwdziałania zagrożeniom, które mogą wpłynąć na system SI.

3.1. ISO/IEC 31000

Norma [9] oferuje kompleksowe i systematycznepodejście do zarządzania ryzykiem, umożliwiając organizacjom proaktywne identyfikowanie, ocenę, postępowanie i monitorowanie potencjalnych zagrożeń. Poprzez wdrożenie ISO/IEC 31000, organizacje mogą ustanowić kulturę zarządzania ryzykiem, co prowadzi do lepszego podejmowania decyzji i skutecznej ochrony aktywów informacyjnych. Jej wdrożenie pozwala na osiągnięcie spójnych i powtarzalnych rezultatów.

3.2. ISO/IEC 27005

Norma [10]to uzupełnienie ISO/IEC 31000, dostarczające wytycznych dla zarządzania ryzykiem bezpieczeństwa informacji. Jej elastyczne podejście, w połączeniu z odpowiednim doświadczeniem, umożliwia częściowe spełnienie wymagań AI Act.

3.3. ISO/IEC 23894

Norma [31] rozwija zasady zarządzania ryzykiem zawarte w ISO 31000, dostarczając praktycznych wskazówek w zakresie systemów AI. Umożliwia dostosowanie procesów do specyficznych potrzeb każdej organizacji, zapewniając bezpieczeństwo i przejrzystość systemów AI.

3.4. ISO/IEC 42001

Norma [8]wspiera organizacje w odpowiedzialnym rozwoju i wykorzystaniu sztucznej inteligencji. Określa wymagania dotyczące zarządzania systemami AI, zapewniając przejrzystość, etykę i ciągłe doskonalenie gwarantując podstawę do zarządzania ryzykami i szansami. Jako jedyna w zestawieniu, umożliwia uzyskanie certyfikatu potwierdzające osiągnięcie wysokiej dojrzałości danej organizacji.

3.5. ISO/IEC 27091

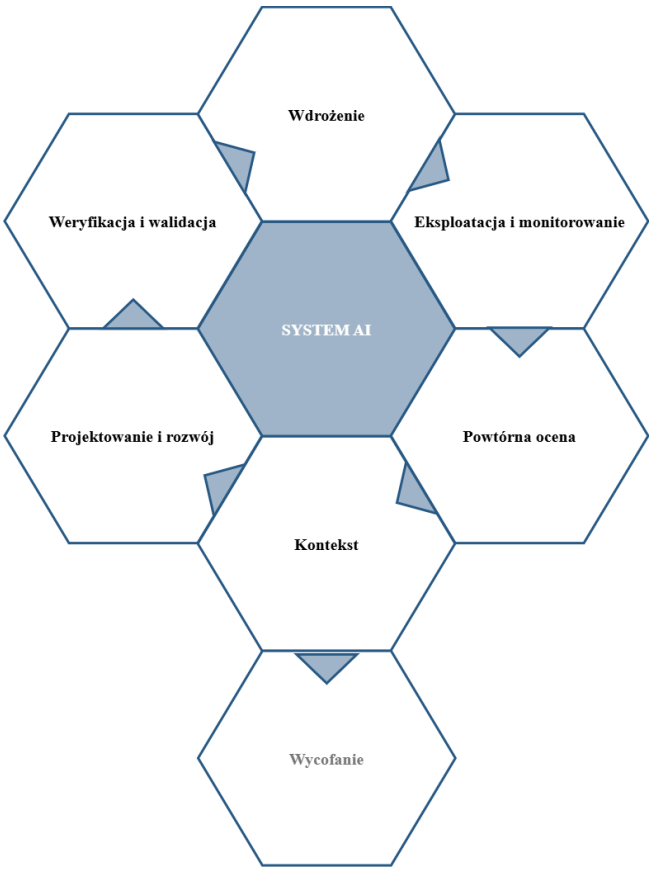
Norma [12] ma stanowić przewodnik dla organizacji, przyczyniając się do identyfikacji i zarządzania ryzykami związanymi z prywatnością danych w systemach AI, promując jednocześnie etyczne i odpowiedzialne praktyki. Planowany termin publikacji koniec 2026 roku.

3.6. Podejście do zarządzania ryzykiem w systemach SI

Wymagania zawarte w akcie [5] oraz wytyczne w normach [8-12, 31] jednoznacznie wskazują na konieczność posiadania wysoko wyspecjalizowanej wiedzy eksperckiej oraz umiejętności praktycznego zastosowania odpowiednich narzędzi i metod, aby sprostać określonym kryteriom w AI Act. Niemniej, warto mieć na uwadze, co też podkreślono w publikacji [32], że pewne niedociągnięcia w rozporządzeniu AI Act mogą prowadzić do nadmiernej regulacji i niepotrzebnych kosztów.Z kolei w pracy [33], czerpano inspirację z metodologii stosowanej w ocenie ryzyka klimatycznego, a także metod stosowanych w prawie, przedstawiono podejście składające się z dwóch etapów: konstruowania scenariuszy wystąpienia ryzyka oraz ilościowej ocenie proporcjonalności. Obie analizowane prace wskazują na istotną rolę konieczności prowadzenia badań empirycznych w obszarze oceny ryzyka związanego z systemami SI.

4.Cykl życia systemu SI

Cykl życia systemu SI to ciągły proces transformacji, począwszy od koncepcji, poprzez rozwój i wdrożenie, aż po ewentualną modernizację lub wycofanie. Chociaż nie istnieje jeden, uniwersalny model tego cyklu, wyróżnia się w nim szereg charakterystycznych etapów i powtarzalnych czynności [34-35]. Procesy te są często iteracyjne, co umożliwia ciągłe doskonalenie systemu i dostosowywanie go do nowych wyzwań i możliwości, jakie niesie ze sobą rozwój technologii SI. Opracowany na podstawie normy ISO/IEC 22989:2022 [4]Rysunek 1 przedstawia uogólniony cykl życia systemu SI.



Rysunek 1.Cykl życia systemu SI (opracowanie własne na podstawie [4]).

Interpretując Rysunek 1 można odnotować następujące procesy w ramach cyklu życia systemu SI:

- Kontekst - na tym etapie precyzowany jest cel systemu, identyfikując potrzeby biznesowe, następnie określane są problemy jakie system SI ma za zadanie rozwiązać wyznaczając możliwe szanse i ryzyka. W następnej kolejności zbiera się informacje o oczekiwaniach wszystkich zainteresowanych stron wraz z oceną wykonalności rozwiązania;
- Projektowanie i rozwój - w ramach tego procesu powstaje architektura systemu SI. Dobierane są odpowiednie technologie i narzędzia, a następnie tworzony jest szczegółowy plan obejmujący m.in. cyklu życia danych [14], opracowanie ogólnej struktury łącząc wszystkie elementy w spójną całość;
- Weryfikacja i walidacja - obejmuje procesy dotyczące zapewnienia jakości poprzez weryfikację funkcjonalności, wydajności, jakości danych i zgodności z wymaganiami;
- Wdrożenie - obejmuje działania od wdrożenia modelu po zarządzanie ryzykiem;
- Eksploatacja i monitorowanie - to procesy związane z monitorowaniem systemu SI, reagowaniem, optymalizacją, aktualizacjami i wsparciem użytkowników;
- Powtórna ocena - proces dotyczy ciągłego doskonalenia systemu AI, z uwzględnieniem zmieniających się wymagań;
- Wycofanie - obejmuje likwidację systemu AI, bezpieczne usuwanie danych oraz w razie potrzeby, utrzymanie starszego systemu.

Cykl życia systemu SI przechodzi przez różne etapy od początkowej koncepcji, przez projektowanie i rozwój, aż po zakończenie użytkowania. Każdy etap jest ważny dla zapewnienia, że system SI spełnia swoje zadanie i przynosi oczekiwane korzyści.

5. Modelowanie zagrożeń w systemach AI

Dynamiczny rozwój systemów opartych na SI oraz ich coraz szersze zastosowanie w różnych dziedzinach życia stwarzają pilną potrzebę wypracowania kompleksowych strategii zarządzania zagrożeniami [22,36-37]. Modelowanie zagrożeń to proces oceny i minimalizacji ryzyka związanego z dowolnym systemem, zwłaszcza tymi wykorzystującymi SI. Działanie to odgrywa znaczącą rolę w analizie bezpieczeństwa systemów SI, pomagając ekspertom w zrozumieniu, w jaki sposób systemy oparte na SI mogą zawieść [21]. W kontekście systemów SI, wiele prac badawczych [15-24] skupia się na dostosowaniu

istniejących metod modelowania zagrożeń do specyfiki tych systemów. W poniższej tabeli 2 zostało przedstawione zestawienie wybranych metod, wraz z ich kluczowymi cechami.

Tabela 2. *Zestawienie wybranych metod modelowania zagrożeń. (opracowanie własne).*

Metoda	Charakterystyka	Zalety	Wady
STRIDE	Skupia się na naruszeniu bezpieczeństwa.	Dojrzała metoda; łatwa do zrozumienia i zastosowania; dobrze udokumentowana; może być stosowana w różnych domenach.	Może nie być wystarczająco szczegółowa dla złożonych systemów; może nie uwzględniać wszystkich typów zagrożeń; ograniczona do analizy bezpieczeństwa, nie uwzględnia aspektów prywatności.
LINDDUN	Skupia się na prywatności.	Komplementarna do STRIDE; specjalnie zaprojektowana do analizy prywatności; może być stosowana w różnych domenach.	Może być trudna do zastosowania w praktyce; wymaga szczegółowej wiedzy na temat systemu i przepisów dotyczących ochrony danych.
PASTA	Skupia się na ryzyku.	Kompleksowe podejście do modelowania zagrożeń; łączy cele biznesowe z wymaganiami technicznymi; bogata dokumentacja	Czasochłonna i pracochłonna; wymaga zaangażowania ekspertów z różnych dziedzin; może być zbyt złożona dla prostych systemów.
VAST	Zautomatyzowane podejście; skupia się na wizualizacji i prostocie	Skalowalna i łatwa do wdrożenia w dużych organizacjach; dostarcza praktycznych i wiarygodnych wyników dla różnych interesariuszy.	Może być czasochłonna; wymaga dużego nakładu pracy przy modelowaniu systemu.

5.1. Metoda STRIDE

W odniesieniu do Tabeli 2, STRIDE to jedna z najpopularniejszych metod modelowania zagrożeń. Została opracowana przez Microsoft w latach 90. Jest stosowana w różnych branżach. STRIDE to akronim oznaczający sześć kategorii zagrożeń:

- Spooofing (podszywanie się): podszywanie się pod inną osobę lub system w celu uzyskania nieautoryzowanego dostępu[17,36].
- Tampering (manipulacja): nieautoryzowana modyfikacja danych lub systemu [19,36].
- Repudiation (kwestionowanie): zaprzeczanie udziałowi w działaniu lub transakcji [19,36].
- Information Disclosure (ujawnienie informacji): nieautoryzowane ujawnienie poufnych danych[17,36].
- Denial of Service (DoS) (odmowa usługi): uniemożliwienie użytkownikom dostępu do systemu lub usługi[19,36].
- Elevation of Privilege (podniesienie uprawnień): uzyskanie wyższych uprawnień w systemie niż te, do których użytkownik jest upoważniony[19,36].

STRIDE jest często używany w połączeniu z diagramami przepływu danych (ang. DFD), które przedstawiają przepływ danych w systemie[19,36]. Analizując DFD, eksperci ds. bezpieczeństwa mogą identyfikować potencjalne zagrożenia STRIDE dla każdego elementu systemu[19].

5.2. Metoda LINDDUN

LINDDUN to metoda modelowania zagrożeń, która koncentruje się na prywatności. Została opracowana w 2015 roku [17]. LINDDUN to akronim oznaczający siedem kategorii zagrożeń [17]:

- Linkability (łączność): możliwość powiązania danych osobowych z daną osobą.
- Identifiability (identyfikowalność): możliwość zidentyfikowania danej osoby na podstawie jej danych osobowych.
- Non-repudiation (niezaprzeczalność): niemożliwość zaprzeczenia udziałowi w działaniu lub transakcji.

- Detectability (wykrywalność): możliwość wykrycia nieautoryzowanego dostępu do danych osobowych lub ich wykorzystania.
- Disclosure of Information (ujawnienie informacji): nieautoryzowane ujawnienie poufnych danych osobowych.
- Unawareness (nieświadomość): brak wiedzy na temat sposobu gromadzenia, przetwarzania i wykorzystywania danych osobowych.
- Non-compliance (niezgodność): naruszenie przepisów dotyczących ochrony danych osobowych.

LINDDUN jest często używany w połączeniu z STRIDE, aby zapewnić kompleksową analizę zagrożeń, uwzględniającą zarówno bezpieczeństwo, jak i prywatność [17].

5.3. Metoda PASTA

PASTA (Process for Attack Simulation and Threat Analysis) to metoda modelowania zagrożeń, która skupia się na ryzyku [17, 22]. Została opracowana przez OWASP (Open Web Application Security Project) i jest szeroko stosowana w modelowaniu zagrożeń dla aplikacji internetowych.

PASTA składa się z siedmiu etapów[17, 22]:

1. Definicji celów: określenie celów biznesowych i technicznych systemu.
2. Definicji zakresu technicznego: określenie komponentów systemu i ich funkcji.
3. Dekompozycji aplikacji: podział systemu na mniejsze komponenty w celu ułatwienia analizy.
4. Analizy zagrożeń: identyfikacja potencjalnych zagrożeń dla każdego komponentu systemu.
5. Analizy podatności i słabości: identyfikacja podatności i słabości w systemie, które mogą zostać wykorzystane przez atakujących.
6. Modelowania ataków: tworzenie modeli ataków, które pokazują, w jaki sposób atakujący mogą wykorzystać podatności i słabości systemu. W tym etapie często wykorzystuje się drzewa ataków.
7. Analizy ryzyka i wpływu: ocena ryzyka i wpływu każdego zagrożenia, biorąc pod uwagę prawdopodobieństwo jego wystąpienia i potencjalne szkody [17,20,22].

5.4. Metoda VAST

VAST (Visual, Agile, and Simple Threat) to zautomatyzowana metoda modelowania zagrożeń, która kładzie nacisk na wizualizację i prostotę. Metoda ta została opracowana z myślą o skalowalności i łatwości wdrożenia w dużych organizacjach. VAST dostarcza praktycznych i wiarygodnych wyników dla różnych organizacji [17]. VAST tworzy dwa modele:

- Model zagrożeń aplikacji, który wykorzystuje diagramy przepływu danych (DFD).
- Model zagrożeń operacyjnych, który wykorzystuje diagramy architektury chmury.

Integracja VAST z cyklem rozwoju oprogramowania i DevOps stanowi istotną zaletę tej metody [17].

5.5. Cykl życia systemu AI, a metody modelowania zagrożeń

Podrozdział prezentuje praktyczne zestawienie przedstawionych metod modelowania zagrożeń, uwzględniając ich zastosowanie w wybranej fazie cyklu życia systemu SI, a także wizualizację tabelaryczną (Tabela 3).

Tabela 3.*Analiza potencjalnych zagrożeń dla systemu SI w wybranym etapie cyklu– połączenie różnych metod. (opracowanie własne).*

Etap cyklu życia systemu SI	Metoda	Zagrożenie	Szczegółowo
Eksploatacja i monitorowanie	STRIDE	Tampering (np. ataki typu evasion)	Ataki mające na celu manipulację danymi wejściowymi modelu.
-	STRIDE	Information Disclosure	Wyciek danych przetwarzanych przez system SI.
-	LINDDUN	Utrudnione wykrycie naruszeń prywatności	Niewłaściwe monitorowanie systemu SI może uniemożliwić wykrycie naruszeń prywatności.
-	PASTA	Nieprawidłowe działanie systemu SI	Brak mechanizmów wykrywania anomalii w działaniu systemu SI może prowadzić do jego nieprawidłowego działania.
-	VAST	Utrudnione wykrycie ataków na system SI	Brak analizy logów systemowych może uniemożliwić wykrycie ataków na system SI.

Tabela 3 prezentuje potencjalne zagrożenia występujące w trakcie eksploatacji i monitorowania systemów SI. Należy zaznaczyć, że specyfika danego systemu może implikować obecność dodatkowych, unikalnych zagrożeń. Z tego względu zaproponowano przyjęcie hybrydowego podejścia do modelowania zagrożeń, które umożliwi kompleksową analizę stanu bezpieczeństwa.

6. Dyskusja

Tradycyjne podejścia do modelowania zagrożeń nie są wystarczające w kontekście złożoności i specyfiki systemów SI, które są podatne zarówno na tradycyjne cyberzagrożenia, jak i na nowe ataki wykorzystujące specyficzne cechy, takie jak ataki typu adversarial machine learning [37].Istnieje luka w obecnych metodach modelowania zagrożeń dla SI, brakuje systematycznych procesów i narzędzi, które skutecznie radzą sobie ze złożonością tych systemów [22]. Aby sprostać temu wyzwaniu niezbędne jest opracowanie specjalistycznego podejścia do modelowania zagrożeń, dopasowanego do specyfiki SI, uwzględniając cały cykl życia systemu SI [22,36].

6.1. Systemowe podejście do zapewnienia bezpieczeństwa systemu SI

Systemowe podejście do zapewnienia bezpieczeństwa systemów SI, uwzględniające regulacje prawne [**5**], zarządzanie ryzykiem [**9**-10,31] oraz modelowanie zagrożeń [15-24], jest niezbędne dla zapewnienia bezpiecznego
i odpowiedzialnego rozwoju i wdrażania SI. Tabela 4 prezentuje, systemowe podejście do modelowania zagrożeń w systemach SI, które łączy wymagania regulacyjne i zarządzanie ryzykiem.

Tabela 4. *Analiza SWOT systemowego podejścia do bezpieczeństwa systemu SI. (opracowanie własne).*

Metryka	Mocne strony	Słabe strony	Szanse	Zagrożenia
Regulacja AI Act	- Wzrost zaufania do systemów SI. - Ochrona praw użytkowników.	- Złożoność regulacji. - Koszty wdrożenia dla organizacji.	- Ujednolicenie rynku SI w UE.	- Opóźnienia we wdrażaniu regulacji. - Brak elastyczności w dostosowywaniu do dynamicznego rozwoju SI.
Zarządzanie ryzykiem	- Systemowe podejście do ryzyka. - Minimalizacja negatywnych skutków.	- Złożoność wdrożenia. - Konieczność ciągłego monitorowania.	- Wzrost świadomości zagrożeń SI. - Rozwój norm/standardów i narzędzi do zarządzania ryzykiem. - Zwiększenie bezpieczeństwa systemów SI.	- Niewystarczająca wiedza i kompetencje.
Modelowanie zagrożeń	- Systemowe podejście do analizy zagrożeń. - Identyfikacja wektorów ataku.	- Ograniczona skuteczność tradycyjnych metod. - Złożoność hybrydowych metod. - Wymaga specjalistycznej wiedzy.	Rozwój zaawansowanych metod modelowania. - Automatyzacja analizy zagrożeń. - Integracja z narzędziami do zarządzania ryzykiem.	- Niewystarczająca ilość danych do analizy.

Kluczowym elementem tego podejścia jest wykorzystanie metod modelowania zagrożeń, stosując podejście hybrydowe, pozwalające na kompleksową analizę specyficznych zagrożeń dla systemów SI. Ważne jest również śledzenie rozwoju regulacji prawnych, takich jak AI Act, oraz dostosowywanie strategii bezpieczeństwa do zmieniających się wymogów.

6.2. Ograniczenia badania

Pomimo kompleksowego podejścia do modelowania zagrożeń zaprezentowanego w tej pracy, uwzględniającego wymagania AI Act [5],należy zdawać sobie sprawę z pewnych ograniczeń.

1. Dobór metody modelowania zagrożeń jest uzależniony od specyfiki danego systemu SI. Zróżnicowanie systemów pod względem architektury, funkcjonalności, danych wejściowych i wyjściowych, a także kontekstu zastosowania, utrudnia opracowanie uniwersalnego podejścia.
2. Modelowanie zagrożeń powinno być integralną częścią całego cyklu życia systemu SI. Należy jednak pamiętać, że różne fazy cyklu życia charakteryzują się

odmiennymi rodzajami zagrożeń i wymagają dostosowania stosowanych metod.
3. Organizacje wdrażające systemy SI mogą napotkać ograniczenia w dostępie do zasobów, m.in. narzędzi, danych czy ekspertów, co może wpłynąć na skuteczność modelowania zagrożeń.
4. Systemy SI rozwijają się w dynamicznym środowisku, w którym stale pojawiają się nowe zagrożenia. Konieczne jest zatem ciągłe monitorowanie i aktualizowanie wiedzy na temat zagrożeń, aby dostosować stosowane metody do zmieniającej się rzeczywistości.
5. Obecnie brakuje ujednoliconego podejścia do modelowania zagrożeń dla systemów SI, co utrudnia porównywanie i ocenę różnych metod. Trwają prace nad normami, takimi jak ISO/IEC 27090, ale ich ostateczny kształt i wpływ na praktykę pozostają otwarte.

7. Podsumowanie

7.1. Wnioski

Praca przedstawia kompleksowe podejście do modelowania zagrożeń na różnych etapach rozwoju systemu SI, spełniając wymagania regulacji AI Act, w zakresie zarządzania ryzykiem. Podkreślono również znaczenie dostępnych narzędzi i metod dla organizacji chcących wdrożyć system SI w sposób zgodny z obowiązującą regulacją. Ponadto, badanie wskazuje, że zaangażowanie ekspertów dziedzinowych na każdym etapie cyklu życia systemu SI jest niezbędne do zapewnienia, że system SI jest skuteczny, ale także bezpieczny dla użytkowników. Dynamiczny rozwój systemów SI sprawia, że modelowanie zagrożeń dla tych systemów jest wciąż obszarem intensywnych badań. Świadczą o tym zarówno proponowane kierunki badawcze, jak i prace nad normą ISO/IEC 27090, która ma na celu opracowanie ogólnych wytycznych i zasad w tym zakresie.

7.2. Kierunki dalszych badań

Fundamentem AI Act jest zasada proporcjonalności, mająca na celu stworzenie takich regulacji, które z jednej strony skutecznie ograniczają potencjalne zagrożenia wynikające z wykorzystania systemów SI, a z drugiej strony nie hamują rozwoju tych technologii i umożliwiają czerpanie z nich pełnych korzyści.Przedstawione podejście do modelowania zagrożeń i zarządzania ryzykiem w systemach SI stanowi podstawę do dalszych badań.

1. Niezbędne jest przeprowadzenie szeroko zakrojonych badań metodami symulacyjnymi i laboratoryjnymi, które pozwolą na ocenę praktycznej użyteczności proponowanych rozwiązań w różnych kontekstach przemysłowych.
2. Konieczne jest opracowanie nowych, bardziej zaawansowanych metod modelowania zagrożeń, które będą w stanie skutecznie radzić sobie ze złożonością współczesnych systemów SI.
3. Przedstawione systemowe podejście wymaga wiedzy eksperckiej i jest czasochłonne w implementacji, monitorowaniu i doskonaleniu, konieczne jest opracowanie rozwiązań automatyzujących.

Literatura

1. Sangwan R., Badr Y., Srinivasan S. Cybersecurity for AI systems: A survey. Journal of Cybersecurity and Privacy, 3(2), 2023, 166-190.
2. Bogdanov D., Etti P., Kamm L., Ostrak A., Pern T., Stomakhin F., Toomsalu M., Valdma S.M., Veldre A. Risks and controls for artificial intelligence and machine learning systems. Version 1.0 [Report]. Estonian Research Institute at Tallinn University of Technology (RIA). 2024. Retrieved November 13, 2024, from https://www.ria.ee/sites/default/files/documents/2024-05/Risks-and-controls-for-artificial-intelligence-and-machine-learning-systems.pdf
3. Knockaert M., Everarts de Velp S., Norouzian M.R., Palacios C., Martínez C., Orduña R., Etxeberria X., Gil A., Pawlicki M., Choras M. (2021). D7.1: AI systems threat analysis mechanisms and tools [Report]. SPARTA project number 830892. Pobrano 13 listopada 2024, z: https://www.sparta.eu/assets/deliverables/SPARTA-D7.1-AI-systems-threat-analysis-mechanisms-and-tools-PU-M18_v1.1.pdf
4. ISO/IEC 22989:2022(E) Information technology — Artificial intelligence — Concepts and terminology [Norma]. International Organization for Standardization, 2022.
5. Parlament Europejski i Rada Artificial Intelligence Act. Pobrano 13 listopada 2024 z:https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=OJ:L_202401689, 2024.
6. Liebl A., Klein T. AI Act: Risk classification of AI systems from a practical perspective [Report]. 2023. Applied AI Initiative. Pobrano 13 listopada 2024, z https://aai.frb.io/assets/files/AI-Act-Risk-Classification-Study-appliedAI-March-2023.pdf
7. Targowski A. Informatyka: modele systemów i rozwoju. Warszawa: Państwowe Wydawnictwo Ekonomiczne, 1980. Pobrano 13 listopada 2024, z https://bcpw.bg.pw.edu.pl/dlibra/doccontent?id=1702

8. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system [Norma]. International Organization for Standardization 2023.
9. ISO/IEC 31000 Risk management — Guidelines [Norma]. International Organization for Standardization 2018.
10. ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements [Norma]. International Organization for Standardization 2022.
11. ISO/IEC 27090 Cybersecurity — Artificial Intelligence — Guidance for addressing security threats to artificial intelligence systems [Norma]. International Organization for Standardization. Pobrano 13 listopada 2024, z:https://www.iso.org/standard/56581.html.
12. ISO/IEC 27091 Cybersecurity and privacy — Artificial Intelligence — Privacy protection [Norma]. International Organization for Standardization.Pobrano 13 listopada 2024,https://www.iso.org/standard/56582.html.
13. ISO/IEC 5338:2023(E) Information technology — Artificial intelligence — AI system life cycle processes [Norma]. International Organization for Standardization 2023.
14. ISO/IEC 8183:2023 Information technology — Artificial intelligence — Data life cycle framework [Norma]. International Organization for Standardization 2023.
15. Pape N., Mansour C. PASTA Threat Modeling for Vehicular Networks Security. In 2024 7th International Conference on Information and Computer Technologies (ICICT) (pp. 474-478). IEEE. https://doi.org/10.1109/ICICT62343.2024.00083, 2024.
16. Stingelová B., Thrakl C.T., Wrońska L., Jedrej-Szymankiewicz S., Khan S., Svetinovic D. User-Centric Security and Privacy Threats in Connected Vehicles: A Threat Modeling Analysis Using STRIDE and LINDDUN. In 2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) (pp. 0690-0697). IEEE 2023 https://doi.org/10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361381
17. Azam N., Michala L., Ansari, S., Truong N. B. Data Privacy Threat Modeling for Autonomous Systems: A Survey From the GDPR's Perspective. IEEE Transactions on Big Data, 2023, 9(2), 388-414.
18. Mauri L., Damiani E. Modeling Threats to AI-ML Systems Using STRIDE. Sensors, 2022, 22(1), 1.
19. Tete S. Threat Modeling and Risk Analysis for Large Language Model (LLM)-Powered Applications. arXiv 2024.
20. von der Assen J., Sharif J., Feng C., Killer C., Bovet G., Stiller B. Asset-Centric Threat Modeling for AI-Based Systems. In 2024 IEEE International Conference on Cyber Security and Resilience (CSR) (pp. 437-444). IEEE 2024.
21. Mauri L., Damiani E. STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets. In 2021 IEEE Cybersecurity Development Conference (CSR) (pp. 147-154). IEEE 2021. https://doi.org/10.1109/CSR51186.2021.9527917
22. Sharif J. Design and Implementation of a Threat Modeling Approach for AI-based Systems (Master's thesis). University of Zurich, Zurich, Switzerland 2023.
23. Tarandach I., Coles M.J. Threat Modeling: A Practical Guide for Development Teams. O'Reilly Media, Inc. 2021.
24. Shostack A. Threat Modeling: Designing for Security. John Wiley & Sons, Inc. 2014.
25. Sportelli M. The AI Act - A Policy Exploration. 2024. DOI: 10.13140/RG.2.2.11397.15847/1.
26. Ehsan U., Riedl M. O. Explainability pitfalls: Beyond dark patterns in explainable AI. Patterns, 2024, 5(6), 100971. https://doi.org/10.1016/j.patter.2024.100971.
27. Simchon A., Edwards M., Lewandowsky S. The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus, 2024, 3(2), pgae035. https://doi.org/10.1093/pnasnexus/pgae035
28. Loefflad C., Grossklags J. How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) (pp. 1515-1530). Association for Computing Machinery, 2024. https://doi.org/10.1145/3630106.3658986
29. Mitka A. The use of “real-time” remote biometric identification systems for law enforcement : comments in light of legislative work on the Artificial Intelligence Act. Problemy Współczesnego Prawa Międzynarodowego Europejskiego I Porównawczego, 2023, 21, 183–202. https://doi.org/10.26106/q3ta-bv90
30. Nair A., Greeshma M. R. Mastering Information Security Compliance Management: A Comprehensive Handbook on ISO/IEC 27001:2022. Packt Publishing Ltd. 2023.
31. ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management [Norma]. International Organization for Standardization 2023.
32. Ebers M. Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU’s AI Act. European Journal of Risk Regulation, 2024 1–20. doi:10.1017/err.2024.78
33. Novelli C., Casolari F., Rotolo A., Taddeo M., Floridi L. AI risk assessment: A scenario-based, proportional methodology for the AI Act. Digital Society, 2024, 3(1), 13. https://doi.org/10.1007/s44206-024-00095-1
34. Muller B., Roth D., Kreimeyer M. Survey of the Role of Domain Experts in Recent AI System Life Cycle Models. In NORDDESIGN 2024 (pp. 256-265).
35. Steidl M., Golendukhina V., Felderer M., Ramler, R. Automation and Development Effort in Continuous AI Development: A Practitioners’ Survey. In 2023 IEEE Symposium on Software Engineering for AI (SEAA) (pp. 120-127). IEEE 2023. https://doi.org/10.1109/SEAA60479.2023.00027
36. Dev J., Akhuseyinoglu N., Kayas G., Rashidi B., Garg V. Building Guardrails in AI Systems with Threat Modeling.

Digital Government: Research and Practice 2024. https://doi.org/10.1145/3674845
37. National Institute of Standards and Technology (NIST). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2024. Pobrano 13 listopada z: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf