AI Security: Twoja przewaga konkurencyjna w wyścigu SI

Przewodnik dla kadry kierowniczej wyższego szczebla: przyspieszanie wdrażania sztucznej inteligencji i zarządzanie ryzykiem

Wprowadzenie

W sytuacji, gdy organizacje pospiesznie wykorzystują transformacyjną moc sztucznej inteligencji, zabezpieczenia często pozostają w tyle za wdrożeniami. Na każdą zatwierdzoną inicjatywę przypada niezliczona liczba przypadków wykorzystania szarej strefy SI przez pracowników i zespoły. Aż 85% osób podejmujących decyzje w obszarze IT twierdzi, że pracownicy wdrażają narzędzia SI zanim zespoły IT zdążą je ocenić.[1] 93% pracowników przyznaje, że wprowadza informacje do narzędzi SI bez uzyskania zgody.[1]

Tymczasem liczba wektorów ataków wykorzystujących sztuczną inteligencję gwałtownie wzrasta — o 47% w ciągu ostatniego roku[2], ponieważ tradycyjne narzędzia zabezpieczeń z trudem dotrzymują kroku zmianom. Nowe formy narażenia bezpieczeństwa danych i luki w zakresie zgodności z przepisami podważają dotychczasowe praktyki w zakresie nadzoru. Liderzy stają więc przed newralgicznym pytaniem: jak zarządzać ryzykiem związanym ze sztuczną inteligencją, nie blokując przy tym innowacji, które umożliwia?

Wymagania są jasne. Organizacje muszą utworzyć środowisko, w którym:

- Wszystkie przypadki użycia SI są znane
- Cała komunikacja SI może być zabezpieczona
- Wszystkie zasady SI mogą być egzekwowane
- Wszystkie modele SI można chronić przed nadużyciami

Pakiet Cloudflare AI Security Suite zapewnia organizacjom poczucie bezpieczeństwa podczas szybkiego wdrażania rozwiązań opartych na sztucznej inteligencji, eliminując niepewność związaną z ryzykiem. Nasza ujednolicona platforma zabezpiecza cały cykl życia sztucznej inteligencji poprzez wykrywanie wykorzystania SI, stosowanie dostępu Zero Trust i ochronę punktów końcowych interfejsu API za pomocą proaktywnych środków obrony przed zagrożeniami. Dzięki zintegrowanemu zarządzaniu danymi zespoły mogą swobodnie wprowadzać innowacje bez uszczerbku dla bezpieczeństwa.

85%
osób podejmujących decyzje w obszarze IT twierdzi, że pracownicy wdrażają narzędzia SI zanim zespoły IT zdążą je ocenić.[1]

Wyzwanie związane z bezpieczeństwem SI

Zagrożenia związane z wdrożeniem SI stymulują rosnący rynek rozwiązań w zakresie bezpieczeństwa tej technologii. Na przykład natywnie chmurowe platformy ochrony aplikacji (CNAPP) koncentrują się na procesach roboczych rozwoju sztucznej inteligencji. Wczesne rozwiązania tego typu stanowią podstawową linię obrony, ale nie wystarczają. Firmy muszą zabezpieczyć cały cykl życia sztucznej inteligencji, w tym chronić systemy SI po ich wdrożeniu do produkcji.

Pełny zakres wyzwań związanych z bezpieczeństwem SI staje się aż nazbyt oczywisty. Obecnie programiści tworzą funkcje SI, pracownicy korzystają z zewnętrznych narzędzi SI, a klienci wchodzą w interakcje z aplikacjami opartymi na SI. Ręczne zabezpieczanie tak różnych środowisk jest skomplikowane, a zachowanie spójności tego procesu – jeszcze trudniejsze.

Zapobieganie niewłaściwemu wykorzystaniu sztucznej inteligencji przez pracowników

Wdrażanie przez pracowników organizacji zatwierdzonych i niezatwierdzonych narzędzi SI wystawia poufne dane i operacje na ryzyko. Zespoły ds. bezpieczeństwa muszą się zmierzyć z następującymi trudnościami:

- **Brak wglądu w wykorzystanie SI:** liderzy często nie mają pełnego obrazu narzędzi SI używanych przez pracowników, miejsc przetwarzania poufnych danych oraz sposobów, w jakie systemy SI łączą się z aplikacjami biznesowymi. To martwe pole stwarza znaczące zagrożenie.
- **Ryzyko związane z bezpieczeństwem danych i zgodnością:** SI zmienia sposób przepływu danych w Twojej organizacji. Dane osobowe, dane stanowiące tajemnicę przedsiębiorstwa oraz dane klientów mogą trafić do systemów sztucznej inteligencji w sposób, który powoduje naruszenia zgodności z przepisami lub ujawnia informacje dające przewagę konkurencyjną.
- **Kontrola dostępu dla przepływów pracy agentowej SI:** zarządzanie dostępem do narzędzi SI dotyczy nie tylko ludzi. Należy również zarządzać dostępem agentowej SI do serwerów MCP i innych systemów o znaczeniu krytycznym. Wymaga to nowego podejścia do zarządzania tożsamością i kontroli dostępu.

Ochrona publicznie dostępnych aplikacji i modeli SI

Niezależnie od tego, czy systemy sztucznej inteligencji są rozwijane wewnętrznie, czy wykorzystywane za pośrednictwem podmiotów trzecich, stają się one kluczowym elementem doświadczeń klientów i użytkowników, dlatego wymagają odpowiedniej ochrony. Raport dotyczący 10 najważniejszych zagrożeń dla modeli LLM według OWASP wskazuje na czynniki zagrożenia, którym nie są w stanie sprostać tradycyjne narzędzia:

- **Ataki z nieograniczonym zużyciem zasobów:** podobnie jak tradycyjne ataki DoS, ataki z nieograniczonym zużyciem zasobów mają na celu przeciążenie modeli LLM żądaniami wymagającymi dużej ilości zasobów. Modele cenowe oparte na płatności za użycie w chmurze mogą powodować gwałtowny wzrost kosztów związanych z takimi atakami, a jednocześnie prawidłowi użytkownicy mogą doświadczać pogorszenia jakości usług.
- **Zatruwanie modeli:** atakujący implementują backdoory, uprzedzenia lub luki w modelach LLM, poprzez wstrzykiwanie uszkodzonych danych do publicznych zbiorów danych lub repozytoriów wykorzystywanych przez programistów do trenowania modeli. Model zatruty w ten sposób zachowuje się normalnie, aż do momentu, gdy określone wyzwalacze aktywują szkodliwe zachowania.
- **Ataki typu „prompt injection”:** Popularna taktyka eksfiltracji danych przez interfejsy SI. Atak typu prompt injection polega na manipulowaniu danymi wejściowymi LLM poprzez osadzanie złośliwych instrukcji w promptach użytkownika lub treściach zewnętrznych, co powoduje, że modele ignorują oryginalne instrukcje i wykonują polecenia atakującego.
- **Jailbreaking:** starannie przygotowane prompty, takie jak scenariusze odgrywania ról, polecenia nadpisujące instrukcje i strategie wieloetapowe, mogą omijać zabezpieczenia LLM i umożliwiać generowanie zabronionych treści lub uzyskiwanie dostępu do poufnych informacji.

Aby umożliwić szybki rozwój innowacji w dziedzinie sztucznej inteligencji bez narażania organizacji na ryzyko, zalecamy liderom przyjęcie kompleksowego podejścia do bezpieczeństwa SI, obejmującego pełne spektrum wymagań:

- Ochrona korzystania z generatywnej SI przez pracowników
- Ochrona aplikacji i obciążeń opartych na SI
- Tworzenie aplikacji opartych na SI z wbudowanymi zabezpieczeniami

Zabezpieczanie przepływów pracy związanych z rozwojem i szkoleniem SI

Projekty SI wykorzystują ogromne zbiory danych, kosztowne zasoby i iteracyjne eksperymenty, co stwarza nowe powierzchnie narażenia na atak i luki w zabezpieczeniach. Zespoły ds. bezpieczeństwa muszą się zająć następującymi kwestiami:

- **Bezpieczeństwo i integralność danych treningowych:** naruszone dane treningowe mogą wprowadzać obciążenia, backdoory lub luki w zabezpieczeniach, które utrzymują się w modelach produkcyjnych. Organizacje muszą zabezpieczyć dostęp do zbiorów danych treningowych, zapobiegać nieautoryzowanym modyfikacjom oraz zapewniać odpowiednie pochodzenie danych w całym cyklu życia modelu.
- **Zarządzanie poświadczeniami i wpisami tajnymi:** przepływy pracy związane z rozwojem sztucznej inteligencji wymagają dostępu do systemów, w tym pamięci masowej w chmurze dla zbiorów danych, klastrów obliczeniowych do trenowania, rejestrów modeli i usług stron trzecich. Nieodpowiednio zabezpieczone hasła lub klucze interfejsu API mogą ujawnić poufne dane uwierzytelniające, umożliwiając nieautoryzowany dostęp do zastrzeżonych modeli, danych treningowych lub systemów produkcyjnych.
- **Kontrola dostępu do środowiska programistycznego:** inżynierowie SI często potrzebują podwyższonych uprawnień, aby eksperymentować z modelami i uzyskiwać dostęp do poufnych danych. Brak odpowiednich mechanizmów kontroli dostępu może prowadzić do zagrożeń wewnętrznych, przypadkowego ujawnienia danych lub nieautoryzowanej ekstrakcji modelu.

Ujednolicenie zabezpieczeń za pomocą rozwiązań Cloudflare

Udane wdrożenie SI koncentruje się na zwiększaniu produktywności, a nie na jej ograniczaniu. Gdy zespoły mogą bez obaw korzystać ze sztucznej inteligencji, mając pewność, że istnieją odpowiednie zabezpieczenia, szybciej wprowadzają innowacje i realizują bardziej ambitne projekty.

Pakiet Cloudflare AI Security Suite tworzy bezpieczne środowisko dla innowacji SI. Ujednolicając możliwości krawędzi usługi bezpiecznego dostępu (SASE) i zabezpieczeń aplikacji internetowych, dyrektorzy mogą łączyć i chronić dwa podstawowe obszary:

- Zewnętrzne, publicznie dostępne aplikacje wykorzystujące sztuczną inteligencję
- Wewnętrzne, prywatne systemy i obciążenia SI

Koncentrując się na całym cyklu życia sztucznej inteligencji, pakiet Cloudflare AI Security Suite zaspokaja potrzeby w zakresie bezpieczeństwa, od wykrywania i zarządzanie ryzykiem po ochronę danych, zabezpieczając dostęp użytkowników oraz chroniąc aplikacje wykorzystujące sztuczną inteligencję i procesy tworzenia oprogramowania. Aby zapewnić kluczową warstwę bezpieczeństwa produkcyjnego w czasie rzeczywistym, globalna sieć Cloudflare działa w trybie inline, sprawdzając i filtrując każdą interakcję SI w celu ochrony danych wszystkich użytkowników i aplikacji.

Zamiast jedynie wykrywać problemy po ich wystąpieniu, Cloudflare zapobiega im i blokuje zagrożenia, zanim dotrą one do modeli SI. W ten sposób zespoły ds. bezpieczeństwa uzyskują wgląd niezbędny do wyprzedzania nowych zagrożeń.

Podstawowe funkcje Cloudflare

Pakiet Cloudflare AI Security Suite zapewnia kompleksowy monitoring, ochronę w czasie rzeczywistym i proaktywne zarządzanie ryzykiem w ramach jednego środowiska, oferując holistyczne podejście do bezpieczeństwa SI.

Kompleksowe możliwości wykrywania i uzyskiwania wglądu w SI

Skuteczność zabezpieczeń SI zależy od możliwości uzyskania w czasie rzeczywistym kompletnego spisu zasobów i przypadków użycia SI — zarówno zatwierdzonych, jak i niezatwierdzonych. U podstaw pakietu Cloudflare AI Security Suite leży ciągłe monitorowanie i automatyczne wykrywanie, które identyfikuje wszystkie modele SI, asystentów, agentów oraz wdrożenia szarej strefy SI we wszystkich typach środowisk — publicznych, prywatnych i wewnętrznych.

Proaktywne zarządzanie ryzykiem SI

Pakiet Cloudflare AI Security Suite pomaga organizacjom zapobiegać atakom poprzez wykrywanie i ograniczanie luk w zabezpieczeniach specyficznych dla sztucznej inteligencji, błędnych konfiguracji i ścieżek ataku, w tym tych z listy 10 najważniejszych zagrożeń dla modeli LLM wskazanych przez organizację OWASP. Ocena wiarygodności aplikacji pomaga priorytetyzować działania naprawcze, umożliwiając zespołom zajęcie się w pierwszej kolejności najpoważniejszymi zagrożeniami.

Bezpieczeństwo aplikacji opartych na SI

Aby pomóc zespołom SecOps na bieżąco śledzić najnowsze wektory zagrożeń SI, pakiet Cloudflare AI Security Suite oferuje proaktywne wykrywanie zagrożeń i mechanizmy ograniczania ryzyka w przypadku specyficznych dla SI luk w zabezpieczeniach, błędnych konfiguracji i ścieżek ataków w potokach SI, w tym ochronę przed atakami typu prompt injection, zatruciem danych i nadużyciami modeli.

Wyspecjalizowane zapory SI wykrywają i etykietują generatywne lub agentowe punkty końcowe interfejsu API, wykrywają próby eksfiltracji danych umożliwiających identyfikację osób i blokują prompty, zanim wpłyną one na wydajność modelu SI lub zatrują model toksyczną zawartością lub dezinformacją.

Dostęp Zero Trust dla przepływów pracy generatywnej i agentowej SI

Zasady Zero Trust, takie jak zasada najmniejszych uprawnień, dotyczą zarówno pracowników, jak i agentów SI. Pakiet Cloudflare AI Security Suite umożliwia egzekwowanie zasad dostępu do sieci w modelu Zero Trust (ZTNA) zarówno dla interakcji człowiek–SI, jak i SI–SI. Dzięki scentralizowanemu rejestrowaniu i mechanizmom kontroli dla serwerów MCP agentowa SI uzyskuje dostęp tylko do zasobów, do których została autoryzowana — w tym przepływów pracy typu just-in-time.

Ochrona danych oparta na sztucznej inteligencji

Skuteczne zabezpieczenia SI wymagają mechanizmu zapobiegania utracie danych, który wykorzystuje różnorodne modele językowe do analizy treści promptów oraz intencji, jakie za nimi stoją. Pakiet Cloudflare AI Security Suite integruje funkcje ochrony przed wyciekiem danych (DLP) w procesach szkoleniowych, promptach i odpowiedziach, aby zapobiegać ujawnianiu danych umożliwiających identyfikację osób, wyciekom danych i nieautoryzowanemu dostępowi w modelach i potokach SI. Wdrożone w trybie inline zabezpieczenia środowiska wykonawczego oparte na interfejsach API stanowią pierwszą linię obrony — szybką i prostą w użyciu — która uzupełnia strategię shift-left zapewnianą przez CNAPP.

Lokalizacja danych

Modele LLM i aplikacje SI podlegają takim samym regulacjom, jak każde inne środowisko danych. Pakiet Cloudflare AI Security Suite pomaga organizacjom zapewnić, że obciążenia SI przestrzegają granic geograficznych i jurysdykcyjnych dzięki zasadom, które utrzymują dane treningowe i żądania wnioskowania w zatwierdzonych regionach.

Wbudowane zabezpieczenia na potrzeby rozwoju SI

Podobnie jak w przypadku każdego oprogramowania, zabezpieczenia aplikacji SI powinny być wbudowywane od samego początku cyklu rozwoju, a nie dodawane później. Pakiet Cloudflare AI Security Suite zapewnia programistom narzędzia i struktury do tworzenia bezpiecznych aplikacji opartych na sztucznej inteligencji już na etapie projektowania.

Wpływ pakietu Cloudflare AI Security Suite na działalność biznesową

Rozwój sztucznej inteligencji to nie tylko ewolucja — to fundamentalna zmiana na miarę industrializacji lub komputeryzacji. Co za tym idzie, szybkie i skuteczne wdrożenie staje się kwestią nie tylko stymulacji wzrostu, ale również zapewnienia organizacji zdolności do przetrwania. Powolne lub niebezpieczne wdrażanie stanowi obecnie zagrożenie egzystencjalne dla organizacji we wszystkich branżach i sektorach.

Pakiet Cloudflare AI Security Suite umożliwia bezpieczne, kontrolowane i efektywne przekształcenia pozwalające sprostać rosnącym oczekiwaniom klientów i wymaganiom rynku w wysoce konkurencyjnych środowiskach.

- **Szybsze innowacje w zakresie sztucznej inteligencji:** gdy zabezpieczenia umożliwiają bezpieczne użytkowanie zamiast blokować wdrożenie, pracownicy i zespoły mogą testować nowe aplikacje SI, aby zwiększyć produktywność w codziennej pracy. Odpowiednie ramy bezpieczeństwa zapewniają programistom solidną podstawę do tworzenia ambitnych funkcji SI bez narażania wrażliwych danych lub systemów.
- **Zmniejszone ryzyko związane ze sztuczną inteligencją:** organizacje mogą zarządzać całym spektrum czynników ryzyka związanych z wdrażaniem sztucznej inteligencji, w tym zagrożeniami nieodłącznie związanymi z aplikacjami internetowymi wykorzystującymi sztuczną inteligencję. Aplikacje SI będące w fazie rozwoju można zabezpieczyć, aby zagwarantować, że pracownicy nie doprowadzą do wycieku poufnych danych lub ich ujawnienia w zbiorach danych treningowych SI. Dział SecOps może proaktywnie identyfikować i minimalizować zagrożenia oraz luki w zabezpieczeniach specyficzne dla SI, zmniejszając powierzchnię narażenia na atak oraz chroniąc kluczowe dane i modele.
- **Usprawnione operacje w zakresie bezpieczeństwa:** scentralizowana widoczność i kontrola nad bezpieczeństwem Twojej SI upraszcza zarządzanie i usprawnia procesy reagowania na incydenty. Twój zespół SecOps może skupić się na strategicznych inicjatywach zamiast nieustannie zajmować się incydentami związanymi ze sztuczną inteligencją.
- **Solidny nadzór nad danymi i zgodność z przepisami:** specyficzne dla SI mechanizmy kontroli w zakresie ochrony danych pomagają zabezpieczyć poufne informacje i dostosować się do zmiennych wymagań regulacyjnych w całym cyklu życia SI.
- **Niższy całkowity koszt posiadania (TCO):** wykorzystanie skonsolidowanej platformy, która integruje istniejące inwestycje w zabezpieczenia, jest bardziej ekonomiczne niż wdrażanie oddzielnych rozwiązań punktowych dla każdego wyzwania związanego z bezpieczeństwem SI.

Typowe przykłady zastosowania pakietu Cloudflare AI Security Suite

Pakiet Cloudflare AI Security Suite doskonale nadaje się do zaspokojenia kluczowych potrzeb związanych z wdrożeniem SI.

- **Zabezpieczenie użytkowania narzędzi SI przez pracowników:** wymuszaj zasady Zero Trust w zakresie dostępu pracowników do publicznych narzędzi generatywnej SI, takich jak ChatGPT, oraz wewnętrznie opracowanych aplikacji opartych na SI.
- **Ochrona publicznie dostępnych aplikacji opartych na SI:** zabezpiecz aplikacje internetowe i interfejsy API, które integrują modele SI — takie jak chatboty i mechanizmy rekomendacji — przed atakami mogącymi zagrozić wrażliwym danym lub doprowadzić do nadużycia modelu.
- **Zarządzanie szarą strefą SI:** automatycznie wykrywaj niezatwierdzone narzędzia SI w całej organizacji i stosuj odpowiednie mechanizmy kontroli w celu umożliwienia dalszych innowacji bez przekraczania zarządzanych granic ryzyka.
- **Oparta na SI ochrona przed wyciekiem danych dla interakcji SI:** zapobiegaj ujawnianiu poufnych danych w promptach lub odpowiedziach SI, zapewniając ochronę informacji poufnych oraz danych umożliwiających identyfikację osób.
- **Bezpieczeństwo rozwoju SI:** zapewnij zespołom inżynieryjnym ramy, dzięki którym bezpieczny rozwój stanie się podejściem domyślnym, umożliwiając im szybkie tworzenie funkcji SI bez uszczerbku dla bezpieczeństwa.

Kwestie dotyczące wdrożenia

Bezpieczeństwo sztucznej inteligencji, będące fundamentalną strategią ochrony, należy traktować z rozwagą.

Zacznij od obecnej infrastruktury	Najskuteczniejsze wdrożenia zabezpieczeń opartych na SI wykorzystują istniejące narzędzia SASE i zabezpieczenia aplikacji, zamiast je zastępować. Takie podejście pozwala oprzeć się na istniejących inwestycjach, jednocześnie rozszerzając zakres ochrony o czynniki ryzyka specyficzne dla sztucznej inteligencji.
Wdróż zunifikowane wbudowane zabezpieczenia	Zdolność do blokowania złośliwej aktywności na brzegu sieci w czasie rzeczywistym ma kluczowe znaczenie dla bezpieczeństwa SI. Wdróż wbudowane mechanizmy kontroli w czasie rzeczywistym i uzupełnij je monitorowaniem opartym na interfejsach API.
Zapewnij pełne pokrycie	Twoja strategia bezpieczeństwa SI powinna uwzględniać pełne spektrum wymagań obejmujące wykorzystanie narzędzi generatywnej SI przez pracowników, ochronę aplikacji i przepływów opartych na SI, zabezpieczenie przepływów agentowej SI oraz bezpieczeństwo przepływów rozwoju SI.
Planuj w skali przedsiębiorstwa	Wybierz rozwiązania, które skalują się wraz z rozwojem Twojego wdrożenia SI. To, co działa w przypadku jednego programu pilotażowego, musi być również możliwe do skalowania w całej organizacji wraz z rosnącym wykorzystaniem SI.
Sprawdź kryteria sukcesu	Sprawdź rozwiązanie w rzeczywistych warunkach przed pełnym zaangażowaniem. Wybierz platformę, która umożliwia bezpłatną, samoobsługową aktywację funkcji korporacyjnych. Dzięki temu możesz przetestować pełny pakiet zabezpieczeń na małą skalę — na przykład w jednym zespole lub aplikacji — aby szybko wykazać jego wartość i upewnić się, że spełnia on kryteria sukcesu.

Następny krok z pakietem Cloudflare AI Security Suite

Wraz z przyspieszeniem wdrożenia SI na całym świecie organizacje, które potrafią bezpiecznie skalować SI, zyskują znaczną przewagę konkurencyjną. Kluczem jest zrozumienie, że bezpieczeństwo SI nie polega na zapobieganiu użyciu SI, ale na jego inteligentnym umożliwieniu.

Pakiet Cloudflare AI Security Suite umożliwia organizacjom wprowadzanie innowacji bez niepewności. Wychodząc od naszych powszechnie stosowanych platform SASE i bezpieczeństwa aplikacji, pakiet Cloudflare AI Security Suite rozszerza ich możliwości, zapewniając zintegrowane wykrywanie SI, kontrolę dostępu Zero Trust, proaktywną obronę przed zagrożeniami opartą na analizie danych oraz niezawodny nadzór nad danymi. Dzięki kompleksowemu zabezpieczeniu modeli i przypadków użycia SI organizacje mogą zapewnić programistom warunki do szybszego tworzenia rozwiązań oraz zwiększyć produktywność pracowników, a wszystko to bez uszczerbku dla komfortu użytkowania przez użytkownika końcowego.

Zaplanuj konsultację

Odkryj, w jaki sposób pakiet Cloudflare AI Security Suite może zrewolucjonizować podejście Twojej organizacji do bezpiecznego wdrażania sztucznej inteligencji.

+44 20 3514 6970

enterprise@cloudflare.com

www.cloudflare.com/pl-pl/

1. ManageEngine. The Shadow AI Surge in Enterprises: Insights from the US & Canada
2. Check Point Research. Q1 2025 Global Cyber Attack Report from Check Point Software: An Almost 50% Surge in Cyber Threats Worldwide, with a Rise of 126% in Ransomware Attacks