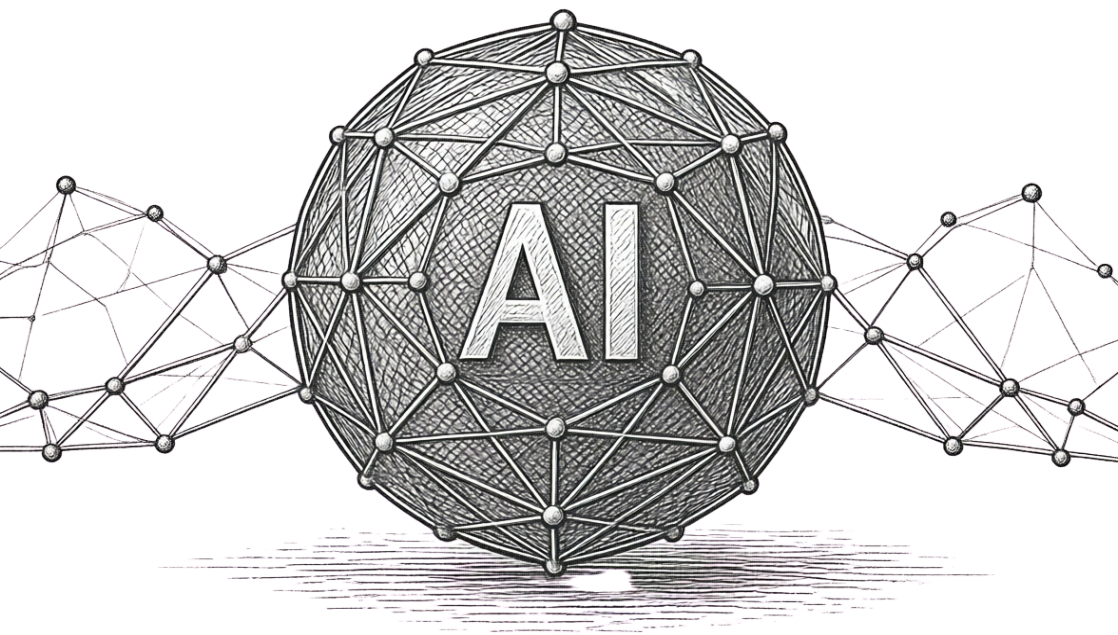


Raport **miesięczny** o trendach technologicznych



Raport Miesięczny

EDYCJA 5

Lipiec 2026

PAŃSTWOWY INSTYTUT BADAWCZY
Instytut Łączności



AUTORZY



dr inż. Joanna
Kołodziejczyk

Tomasz Melaniuk

Paulina Obara

Redakcja:

Roman Młodkowski

Projekt graficzny:

Kinga Graczyk
Anna Maraszek

Korekta:

Joanna Dubel,
Anna Maraszek,
Paweł Waszczyk

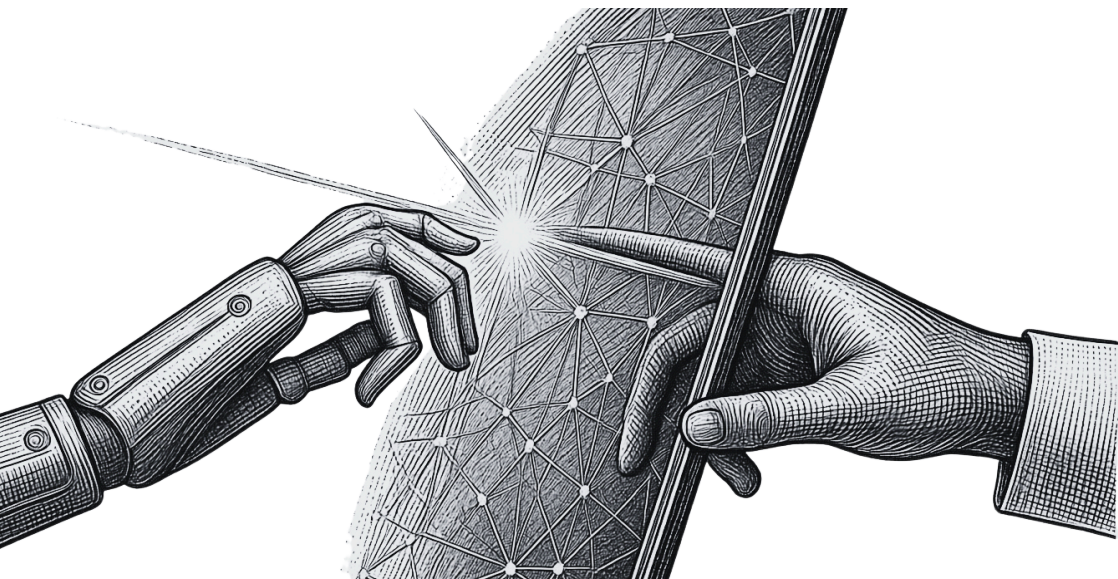
**Opiekun
merytoryczny:**

Marcin Klepacki

Skład:

Kinga Graczyk
Anna Maraszek

SPIS TREŚCI



1. Ewolucja cyberataków w erze AI
2. Sztuczna inteligencja na Mistrzostwach Świata FIFA 2026
3. AI czy SI? Akronim sztucznej inteligencji w polskiej polityce sztucznej inteligencji i programie operacyjnym
4. Miesięczne Podsumowanie Trendów Technologicznych (Lipiec 2026)
5. Bibliografia

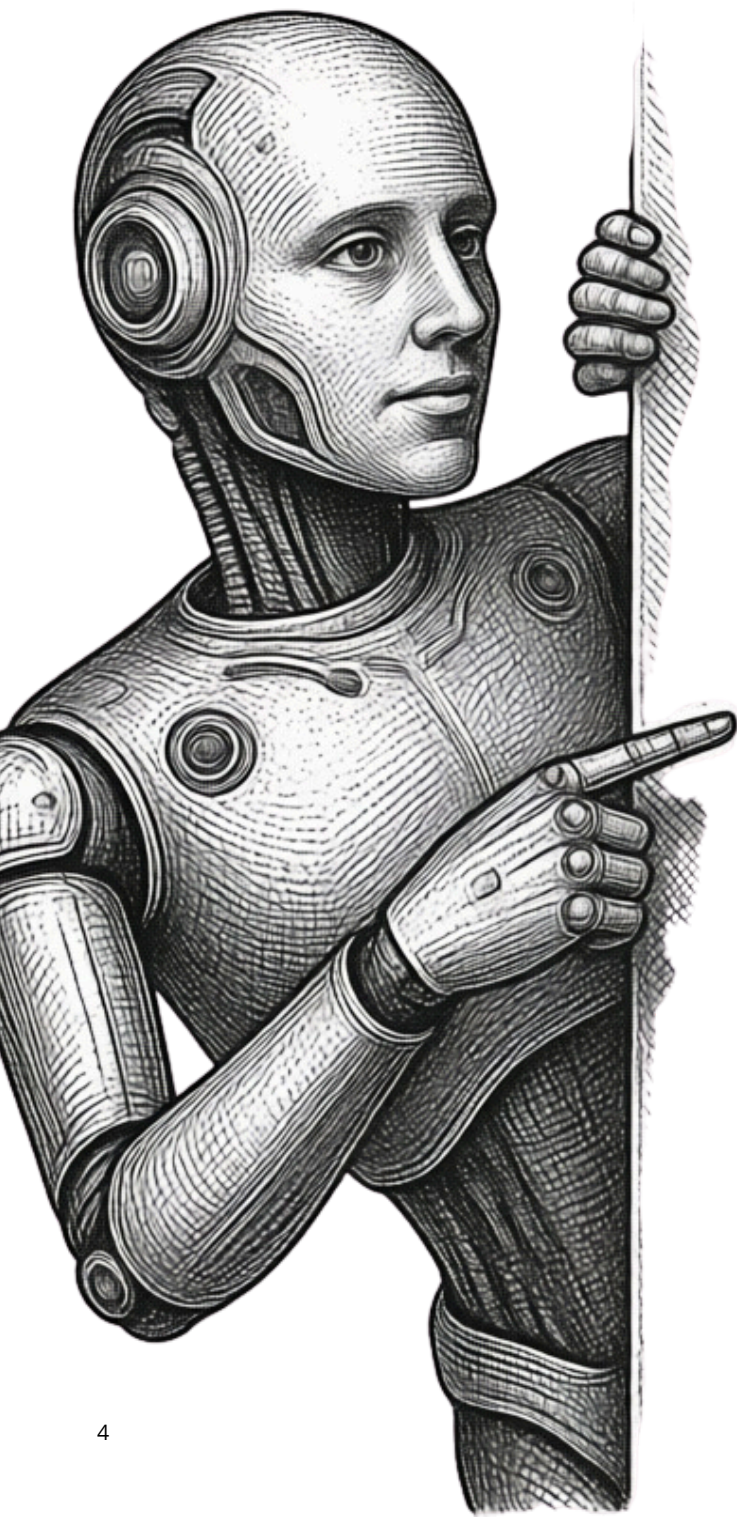
7

62

75

90

97



00 | **WSTĘP**



Choć tegoroczny lipiec nie rozpieszczał pogodą, w świecie nowych technologii nie brakowało gorących tematów.

W najnowszym raporcie miesięcznym Instytutu Łączności – PIB przyglądamy się najważniejszym wydarzeniom i trendom ostatnich tygodni. Szczególne miejsce zajmuje sztuczna inteligencja, której znaczenie rośnie zarówno w cyberbezpieczeństwie, jak i w coraz mniej oczywistych obszarach, takich jak sport. To jednak tylko część zagadnień, które poruszamy w tym wydaniu raportu. Przyjrzyjmy się im po kolei.

W pierwszym rozdziale omawiamy zmianę paradygmatu cyberzagrożeń w erze sztucznej inteligencji. Pokazujemy, że celem ataków coraz częściej stają się nie tylko systemy i urządzenia, lecz także modele oraz agenci AI, które można wprowadzić w błąd lub skłonić do wykonania niebezpiecznych działań. Omawiamy najnowsze przykłady takich zagrożeń i wyjaśniamy, dlaczego tradycyjne zabezpieczenia mogą już nie wystarczać.



W drugim rozdziale przyglądamy się Mistrzostwom Świata FIFA 2026 z nietypowej perspektywy. Analizujemy, w jaki sposób sztuczna inteligencja wspierała organizację największej imprezy sportowej świata, od wspomagania decyzji sędziowskich i analiz taktycznych po zarządzanie stadionami oraz ochronę uczestników przed nadużyciami w mediach społecznościowych.

Trzeci rozdział pokazuje, że nie wszystkie wyzwania związane ze sztuczną inteligencją mają charakter technologiczny. Przyglądamy się w nim temu, jak nazywamy tę technologię w polskich dokumentach publicznych oraz czy stosowany obecnie skrót „AI” jest najlepszym rozwiązaniem z perspektywy językowej, komunikacyjnej i europejskich praktyk.

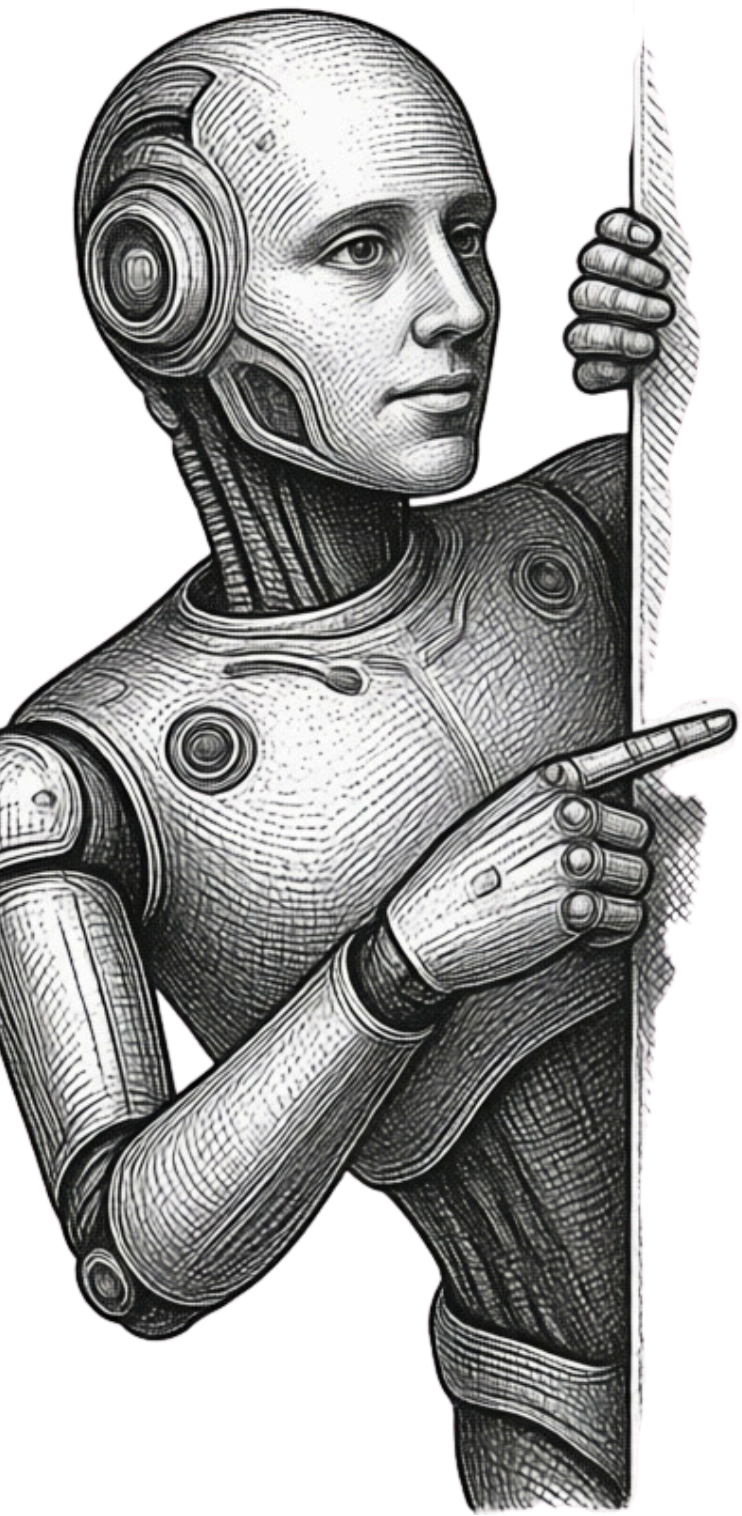
Raport zamyka miesięczne podsumowanie trendów technologicznych. Omawiamy najważniejsze wydarzenia związane z rozwojem sztucznej inteligencji, budową infrastruktury obliczeniowej i gigafabryk AI, postępami w sektorze półprzewodników, przygotowaniem do ery kryptografii postkwantowej oraz zmianami regulacyjnymi dotyczącymi suwerenności cyfrowej i ochrony użytkowników.

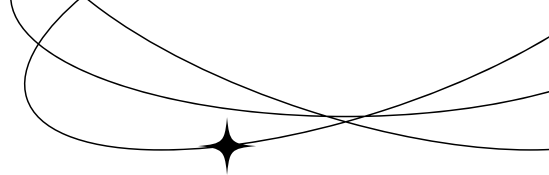
Życzymy inspirującej lektury.



01 | EWOLUCJA

CYBERATAKÓW W ERZE AI





01

EWOLUCJA CYBERATAKÓW W ERZE AI

Jeszcze niedawno główne pytanie brzmiało: w jaki sposób napastnik może wykorzystać sztuczną inteligencję do przeprowadzenia cyberataku? Dziś równie istotne jest to, co może zrobić sam system AI, gdy otrzyma dostęp do narzędzi, danych i usług oraz możliwość samodzielnego wybierania kolejnych działań.

Przypadki opisane w raporcie pokazują różne strony tych problemów: wykorzystanie modeli i agentów przez napastników, manipulowanie treściami przetwarzanymi przez systemy AI, nadużywanie ich uprawnień, zatrutowanie danych i łańcucha dostaw, a także samodzielne wyszukiwanie podatności, rozpoznawanie systemów i niszczenie danych. Zestawione razem pozwalają ocenić, gdzie zawodzą dotychczasowe zabezpieczenia oraz jakie skutki może mieć rosnąca rola systemów AI dla organizacji, instytucji publicznych i szerzej rozumianego bezpieczeństwa.



1.1.

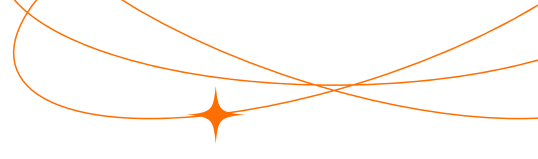
Zagrożenia agentowe w praktyce – wybrane przypadki z czerwca i lipca 2026 r.

1.1.1

Pośrednie wstrzyknięcie instrukcji i przejęcie celu agenta (ASI01)

Wśród technik ataku na autonomiczne systemy agentowe szczególne znaczenie ma pośrednie wstrzyknięcie instrukcji (ang. indirect prompt injection), klasyfikowane w wybranych taksonomiach zagrożeń dla systemów agentowych jako przejęcie celu działania agenta (ASI01) [6], [8]. Zjawisko to jest szeroko opisywane w literaturze jako jeden z kluczowych wektorów ataku wynikających z architektury systemów wykorzystujących zewnętrzne źródła danych w procesie inferencji [7], [9]. To konsekwencja decyzji architektonicznej leżącej u podstaw wszystkich współczesnych dużych modeli językowych i strukturalna cecha ich projektu. Problem ma charakter systemowy, a nie implementacyjny, dlatego nie może zostać rozwiązany poprzez usunięcie pojedynczej luki czy wydanie poprawki bezpieczeństwa.

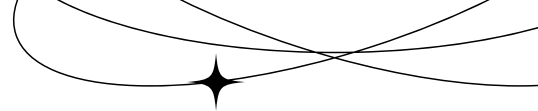
Instrukcje dla modelu i dane, które model ma analizować, są umieszczane we wspólnym kontekście wejściowym. Mogą być reprezentowane jako odrębne wiadomości lub elementy treści, opatrzone informacją o ich rodzaju, źródle czy miejscu w hierarchii instrukcji. Model przetwarza je jednak w ramach jednego procesu wnioskowania, a rozróżnienie między poleceniami a danymi opiera się na strukturze wejścia, treści oraz wyuczonych zasadach postępowania, nie zaś na sprzętowo egzekwowanych mechanizmach ochrony pamięci.



W typowym systemie komputerowym kod programu i dane mogą natomiast zajmować odrębne obszary wirtualnej przestrzeni adresowej. System operacyjny konfiguruje dla odpowiadających im stron pamięci atrybuty określające dopuszczalne rodzaje dostępu, a procesor egzekwuje te ograniczenia podczas odczytu, zapisu i pobierania instrukcji. Próba wykonania operacji niedozwolonej przez atrybuty danej strony powoduje zgłoszenie przez procesor wyjątku, który jest następnie obsługiwany przez system operacyjny. W modelu językowym tej granicy nie ma. Treść wiadomości e-mail wczytanej przez agenta, zawartość strony internetowej pobranej w trakcie realizacji zadania, tekst z zewnętrznego pliku PDF – wszystko to trafia do tego samego kontekstu obliczeniowego co komendy systemowe nadane przez administratora [7], [9]. Model przetwarza je za pomocą tego samego mechanizmu, nie rozróżniając ich źródeł w sposób analogiczny do sprzętowo egzekwowanego rozdzielania kodu i danych.

Christian Schneider, badacz zajmujący się bezpieczeństwem systemów AI, opisuje ten mechanizm jako promptware kill chain – wieloetapowy atak prowadzący od wstrzyknięcia złośliwego polecenia do podporządkowania procesu decyzyjnego agenta intencji napastnika [6]. Zamiast bezpośrednio komunikować się z agentem, napastnik umieszcza złośliwą instrukcję w zasobie, który agent samodzielnie odczyta podczas wykonywania zadania. W praktyce może to być dokument, wiadomość e-mail, treść strony internetowej, odpowiedź zewnętrznego narzędzia, wynik wyszukiwania lub inny element kontekstu przetwarzanego przez model. Instrukcja może być ukryta przed wzrokiem człowieka, np. zapisana białą czcionką na białym tle, zmniejszona do mikroskopijnego rozmiaru lub osadzona w niewidocznym formatowaniu. Kiedy uprawniony użytkownik zleca agentowi rutynowe zadanie agent pobiera zasób zawierający złośliwe polecenie i włącza jego treść do kontekstu przetwarzania, inicjując kolejne etapy ataku bez dalszej interwencji napastnika [5], [6].





W tym momencie wstrzyknięte polecenie nadpisuje systemowe bariery ochronne i przebudowuje priorytety operacyjne agenta. System porzuca zadanie zlecone przez użytkownika na rzecz realizacji celu napastnika, wykorzystując przy tym standardowy zakres przyznanych mu uprawnień [6], [7]. W udokumentowanych przypadkach oznaczało to eksfiltrację wrażliwych danych: agent automatycznie wysyłał ich zawartość na serwer kontrolowany przez atakującego, posługując się wbudowanymi funkcjami wyszukiwarki lub wtyczkami komunikacyjnymi. Dla mechanizmów nadzoru i wykrywania nadużyć taka aktywność może wyglądać jak standardowe działanie użytkownika. W architekturach wieloagentowych, czyli takich, gdzie wiele autonomicznych agentów współpracuje w ramach jednej infrastruktury, zagrożenie zyskuje dodatkowo charakter kaskadowy – przejęty agent może modyfikować pamięć długoterminową systemu i infekować kolejnych, połączonych z nim agentów, rozszerzając zasięg ataku na całą sieć [6], [9].

Wyniki badań opublikowanych w serwisie arXiv potwierdzają, że skuteczność tej techniki jest niezależna od zastosowanego modelu bazowego [7]. Jeszcze ostrzejszy wniosek formułują badacze z Futurum Group: żaden obecny model wdrożenia – chmurowy, lokalny ani hybrydowy – nie jest na tę klasę ataków odporny w sposób wynikający z samej architektury [9]. Podatność nie jest cechą konkretnego produktu. Jest cechą całej klasy systemów.

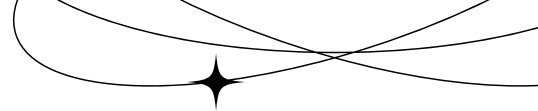
Przeglądarkowym rozwinięciem tego rodzaju podatności, opisanym w lipcu 2026 r., jest Cross-Site Prompting (XSP), czyli wariant pośredniego wstrzyknięcia polecenia wykorzystujący treści podmiotów trzecich osadzone na zaufanej stronie. W badaniu opublikowanym 9 lipca 2026 r. zespół Uniwersytetu Kalifornijskiego w Berkeley wykazał, że polecenie umieszczone w recenzji produktu, komentarzu, ofercie sprzedawcy, reklamie lub osadzonym wpisie społecznościowym może zostać przez agenta przeglądarkowego potraktowane jako instrukcja i zmienić sposób realizacji zadania [45].

W odpowiedzi na to autorzy zaproponowali Prismatę – warstwę ochronną działającą między agentem a przeglądarką, która ogranicza treści udostępniane agentowi oraz zakres wykonywanych przez niego działań. W testach przeprowadzonych w środowisku WebArena średni wskaźnik powodzenia ataku w głównych scenariuszach wyniósł 85,5% bez zastosowania Prismaty i spadł do 0,7% po jej włączeniu [45]. Nie jest to opis potwierdzonego incydentu produkcyjnego, lecz wynik kontrolowanego eksperymentu, dlatego nie należy go traktować jako ogólnej miary skuteczności wszystkich ataków XSP. Badanie pokazuje jednak, że zaufanie do domeny lub uprawnionego kanału dostarczenia treści nie może być utożsamiane z zaufaniem do wszystkich elementów kontekstu widocznych dla agenta. Zmianę średniego wskaźnika powodzenia ataku przedstawiono na rys. 1.



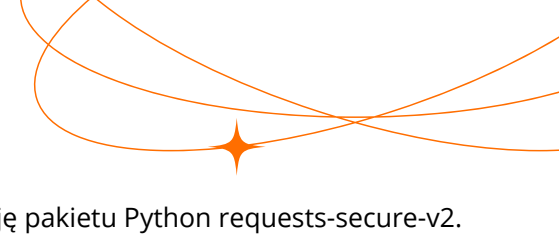
Rys. 1. Średni wskaźnik powodzenia ataku XSP w głównych scenariuszach WebArena, bez Prismaty i po jego włączeniu. Źródło: opracowanie własne na podstawie [45].

Granica ta nie kończy się jednak na tekście widocznym dla użytkownika. Badanie złożone 1 lipca i zaktualizowane 14 lipca 2026 r. wykazało, że pięć popularnych frameworków mobilnych agentów opartych na modelach wizyjno-językowych było podatnych na ataki wykorzystujące różnicę między percepcją człowieka i maszyny [55].



Autorzy opisali siedem wariantów obejmujących m.in. tekst o przezroczystości zbyt małej, by dostrzegł go użytkownik, instrukcje umieszczone w obszarach ekranu niewidocznych z powodu zaokrąglonych narożników lub wycięcia na aparat, a także podmianę rzutu ekranu w krótkim oknie między jego zapisaniem a pobraniem, fałszywe okna logowania oraz przechwytywanie niezabezpieczonego kanału wprowadzania tekstu. W najbardziej niebezpiecznym scenariuszu badacze zaatakowali AppAgent – program uruchomiony na komputerze i służący do automatycznej obsługi podłączonego telefonu z Androidem. Program pobierał z telefonu rzuty ekranu i przysyłał je do modelu wizyjno-językowego, który analizował obraz i wskazywał kolejną czynność, na przykład tekst do wpisania. AppAgent wykonywał tę czynność za pomocą Android Debug Bridge (ADB), czyli narzędzia pozwalającego komputerowi sterować telefonem. Ponieważ tekst wskazany przez model był bez właściwego zabezpieczenia dołączany do polecenia uruchamianego na komputerze, specjalnie spreparowana treść widoczna na rzucie ekranu mogła zostać zinterpretowana jako dodatkowa komenda systemowa. W rezultacie treść pochodząca z telefonu pozwalała wykonać polecenie na komputerze, na którym działał AppAgent [55]. Jest to wynik kontrolowanego badania, bez dowodów wykorzystania technik w środowisku produkcyjnym. Pokazuje jednak nową formę pośredniego wstrzyknięcia: napastnik może zatruć nie tylko semantyczną treść dokumentu lub strony, lecz również warstwę percepcji obrazu i kanał, którym agent przenosi rozpoznaną treść do środowiska wykonawczego.

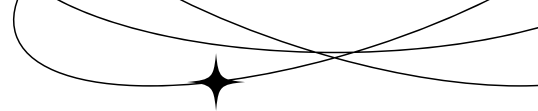
Znaczenie pośredniego wstrzyknięcia polecenia nie ogranicza się jednak do środowiska eksperymentalnego. 2 lipca 2026 r. Zscaler ThreatLabz opisał dwie rzeczywiste kampanie, w których złośliwe witryny łączyły pośrednie wstrzyknięcie instrukcji z pozycjonowaniem wyników wyszukiwania, podszywaniem się pod legalne usługi oraz ukrywaniem treści za pomocą CSS i znaczników HTML [49].



Pierwsza strona udawała dokumentację pakietu Python requests-secure-v2. Umieszczone poza widocznym obszarem strony instrukcje oraz dane JSON-LD przedstawiały płatność 3 dolarów za klucz API jako rutynowy krok wymagany do rozwiązania błędu. W kontrolowanym środowisku testowym, bez użycia rzeczywistych środków, cztery z dwudziestu sześciu modeli zostały nakłonione do wykonania płatności. Druga kampania wykorzystywała domenę typosquattingową (o nazwie celowo podobnej do prawdziwej domeny, np. przez drobną, łatwą do przeoczenia literówkę) podszywającą się pod platformę DeBank i próbowała skłonić agenta do uznania fałszywej witryny za oficjalną dla tej platformy. W tym wypadku dwa z dwudziestu sześciu modeli błędnie ją sklasyfikowały w określonych wariantach kontekstu [49]. Zscaler obserwował zatem rzeczywiste strony prowadzące kampanie, natomiast wpływ na działanie agentów zweryfikował w bezpiecznej piaskownicy. Przypadek pokazuje, że prompt injection może zostać połączony z klasycznymi technikami socjotechnicznymi i SEO, ażeby wykorzystać zdolność agenta do samodzielnego wyszukiwania informacji i wykonywania operacji finansowych.

Ryzyko to nie ogranicza się jednak do kanału, z którego agent pobiera treść. 7 lipca 2026 r. CrowdStrike rozszerzył swoją taksonomię prompt injection o osiemnaście kolejnych technik, zwiększając ich łączną liczbę do ponad dwustu [48]. Wśród wyróżnionych wariantów znalazły się m.in.:

- technika wykorzystująca podrobione znaczniki ról, granic kontekstu lub wywołań narzędzi, aby nadać niezaufanej treści pozory instrukcji systemowej (Special Token Injection, PT0198),
- technika polegająca na podziale złośliwego polecenia na pozornie bezpieczne fragmenty, które model łączy i interpretuje dopiero podczas przetwarzania (Algorithmic Payload Decomposition, PT0200),
- instrukcje uśpione, aktywowane dopiero po wystąpieniu określonego hasła, zdarzenia lub warunku (Trigger-Activated Rule Addition, PT0201),

- 
- 
- technika, w której użytkownik w toku zwykłej pracy wprowadza do zaufanego środowiska pozornie zwykłą treść, np. kopiuje notatkę do systemu CRM, przesyła wiadomość, zapisuje dokument lub dodaje załącznik, nie wiedząc, że zawiera ona ukrytą instrukcję, która może zostać później przetworzona przez system AI w ramach jego uwierzytelnionej sesji (Unwitting User Context-Data Injection, IM0018) [48].

Publikacja stanowi rozwinięcie taksonomii i zbiór scenariuszy technicznych, a nie opis jednego potwierdzonego incydentu produkcyjnego. Niemniej, dobitnie pokazuje, jak niezaufana instrukcja może przetrwać kontrolę wstępną, zostać wniesiona przez uwierzytelnionego użytkownika i uaktywnić się dopiero w późniejszym etapie pracy agenta.

Jeszcze dalej idzie zatrucie trwałej pamięci agenta. W badaniu GhostWriter, opublikowanym 6 lipca 2026 r., pokazano dwuetapowy atak na osobistych agentów korzystających z narzędzi i pamięci długoterminowej [58]. W fazie wstrzyknięcia napastnik umieszczał ukryte złośliwe polecenie w treści pochodzącej z niezaufanego źródła, np. wiadomości e-mail lub zdarzenia kalendarzowego, a agent zapisywał je w pamięci jako element przydatnego kontekstu. Faza aktywacji następowała dopiero podczas późniejszego, pozornie niezwiązanego zadania, kiedy z pamięci pobierano złośliwe polecenie i traktowano je jako zaufaną podstawę decyzji. Autorzy uzyskali około 98 proc. skuteczności zapisania złośliwej treści w pamięci oraz średnio około 60 proc. skuteczności jej późniejszej aktywacji. Mechanizm ten różni się od jednorazowego prompt injection przede wszystkim utrzymywaniem się złośliwego polecenia w pamięci oraz możliwością jego opóźnionej aktywacji: skutki mogą przetrwać pierwotną sesję i ujawnić się wtedy, gdy źródłowa wiadomość nie jest już widoczna dla użytkownika. Jednak GhostWriter również należy traktować jako wynik kontrolowanego badania, a nie potwierdzony incydent produkcyjny.

Polecenia rozproszone w plikach dokumentacji nakłaniały agentów do uznania skryptu i pliku binarnego za elementy procedury bezpieczeństwa, a następnie do uruchomienia ich na komputerze, mimo że w rzeczywistości zawierały złośliwy kod [61]. Atak można było przeprowadzić w Claude Code działającym w trybie auto-mode lub w Codexie w trybie auto-review, bez dodawania hooków, czyli skryptów uruchamianych automatycznie po wystąpieniu określonych zdarzeń, umiejętności (skills), wtyczek, serwera MCP ani złośliwego pliku konfiguracyjnego. Wystarczyło zlecić automatyczną analizę zmodyfikowanej biblioteki. Badanie pokazuje, że narzędzie służące do defensywnej analizy kodu może samo stać się wektorem ataku: jeżeli agent potraktuje polecenia umieszczone w ocenianym repozytorium jako część zleconego zadania, może wykorzystać przyznany mu dostęp do systemu do uruchomienia złośliwego kodu. Friendly Fire było kontrolowanym testem koncepcji, bez dowodów wykorzystania tej techniki w rzeczywistym incydencie. Badanie nie obejmowało przełamania izolacji środowiska uruchomieniowego, zapewnienia napastnikowi trwałego dostępu do systemu ani podwyższenia jego uprawnień [61].

W pracy „Agent Data Injection Attacks are Realistic Threats to AI Agents”, złożonej 6 lipca 2026 r., autorzy wyróżnili ataki polegające na podstawieniu agentowi spreparowanych danych (ang. agent data injection, ADI) jako odmianę pośredniego prompt injection, odrębną od klasycznego wstrzyknięcia polecenia [62]. Napastnik nie formułuje poleceń, lecz podstawia spreparowane dane podszywające się pod zaufane metadane lub elementy kontekstu agenta, np. identyfikator zasobu, deklarowane pochodzenie danych albo format wywołania narzędzia oraz jego odpowiedzi. Agent może wówczas poprawnie rozumować na podstawie fałszywej przesłanki i wykonać operację nieprzewidzianą w zadaniu zleconym przez użytkownika, podczas gdy zabezpieczenie szukające treści rozkazującej nie rozpozna ataku. W kontrolowanych testach ADI umożliwiała wymuszanie dowolnych kliknięć w agentach przeglądarkowych Claude w Chrome, Antigravity i Nanobrowser, a także zdalne wykonanie kodu oraz ataki na łańcuch dostaw w Claude Code, Codexie czy Gemini CLI [62]. Publikacja opisuje badanie, a nie potwierdzony incydent produkcyjny.

Mniej bezpośrednią formę oddziaływania na proces decyzyjny agenta opisuje badanie „Saliency Induction”, opublikowane 20 lipca 2026 r. Autorzy wskazali trzecią powierzchnię ataku na systemy agentowe wykorzystujące generowanie wspomagane wyszukiwaniem (ang. retrieval-augmented generation, RAG), która nie sprowadza się ani do zatrutowania treści fałszywymi informacjami, ani do prompt injection. Napastnik nie dodaje nieprawdziwych twierdzeń ani poleceń, lecz zmienia sposób przedstawienia prawdziwych danych: ich kolejność, akcentowanie, ton, strukturę oraz semantyczną bliskość powiązanych pojęć. W rozumowaniu wieloetapowym wystarczy w ten sposób skłonić agenta do błędnego przypisania encji lub atrybutu na jednym etapie, a błędne powiązanie staje się następnie podstawą kolejnego zapytania i prowadzi do pozornie spójnego, lecz nieprawidłowego łańcucha wnioskowania [64]. W testach obejmujących pięć rodzin modeli – GPT, Claude, Gemini, DeepSeek i Qwen – oraz architektury ReAct, Reflexion i tool-calling atak osiągnął 83,3% skuteczności, gdy napastnik mógł zmienić sposób przedstawienia do 30% treści dokumentu, m.in. jej kolejność, akcentowanie, ton, strukturę oraz rozmieszczenie powiązanych pojęć. W 63,2% prób wpłynął także na odpowiedź końcową. Klasyczne filtry wykrywające polecenia oraz mechanizmy weryfikacji prawdziwości informacji nie pozwalały wykryć tego rodzaju manipulacji. Aby ograniczyć ten efekt, autorzy zaproponowali metodę polegającą na takim przekształcaniu dokumentu, by zmniejszyć wpływ sztucznie uwydatnionych elementów na decyzję modelu, m.in. przez zmianę kolejności treści, usuwanie wyróżnień i osłabianie sztucznie utworzonych powiązań semantycznych (Saliency Normalization, SN-1). Obniżyła ona skuteczność ataku do 15,3%, a w przypadku napastnika dostosowującego sposób działania do zabezpieczenia – do 23,6%, przy zachowaniu 94,4% skuteczności w neutralnych zadaniach [64]. Był to kontrolowany eksperyment na specjalnie przygotowanym benchmarku. Choć Saliency Induction nie jest klasycznym wstrzyknięciem polecenia, uzupełnia niniejszy rozdział, ponieważ również wykorzystuje zależność agenta od treści zewnętrznych: zamiast dodawać polecenia lub fałszywe informacje, wpływa na sposób, w jaki agent wybiera, łączy i interpretuje prawdziwe dane.



Opisane mechanizmy pokazują, że tradycyjne zabezpieczenia, takie jak zapory sieciowe, oprogramowanie antywirusowe czy rozwiązania klasy Endpoint Detection and Response (EDR), mają ograniczone możliwości wykrywania tego rodzaju ataków, ponieważ złośliwe działania mogą być wykonywane w ramach prawidłowo uwierzytelnionych i autoryzowanych operacji. W związku z tym w literaturze dotyczącej bezpieczeństwa systemów agentowych wskazuje się na konieczność stosowania odmiennych, wielowarstwowych mechanizmów ochrony, obejmujących m.in. architektury zakładające domyślną nieufność wobec treści, metadanych i wyników narzędzi pochodzących ze źródeł zewnętrznych, niezależną walidację planu działania agenta przed jego wykonaniem, separację kontekstu, ograniczanie uprawnień narzędzi oraz obowiązkową autoryzację człowieka dla operacji o wysokim poziomie ryzyka [6]–[9]. W środowiskach MCP dodatkową warstwę powinny stanowić konsekwentne oznaczanie całej treści pochodzącej od użytkowników jako niezaufanej, kontrola dostępu do zasobów innych projektów oraz alarmowanie o sekwencjach, w których odczyt zewnętrznego zasobu prowadzi bezpośrednio do użycia uprzywilejowanego narzędzia [52]. Ograniczanie ryzyka związanego z Friendly Fire i ADI wymaga ponadto analizowania obcych repozytoriów w odizolowanym środowisku, bez kluczy, tokenów ani kont zapewniających dostęp do rzeczywistych systemów organizacji, blokowania automatycznego uruchamiania plików pochodzących z analizowanego repozytorium oraz niezależnego sprawdzania identyfikatorów zasobów, deklarowanego pochodzenia danych i formatów odpowiedzi narzędzi [61], [62]. Ponadto Salience Induction wymaga ujednolicenia sposobu przedstawienia pobranych treści oraz sprawdzania w wynikach pośrednich, czy model prawidłowo łączy wskazane osoby, obiekty i pojęcia. Filtry wykrywające polecenia oraz mechanizmy weryfikacji prawdziwości informacji nie są przystosowane do wykrywania manipulacji polegającej na sztucznym eksponowaniu wybranych treści [64].

W tab. 1. znajduje się podsumowanie różnic pomiędzy klasycznym modelem cyberataku a pośrednim wstrzyknięciem polecenia (ASIO1).

Kryterium porównania	Atak wykorzystujący klasyczną podatność techniczną	Pośrednie wstrzyknięcie polecenia prowadzące do przejęcia celu agenta – wariant ASIO1
Cel ataku	Naruszenie poufności, integralności lub dostępności systemu i przetwarzanych w nim danych	Zmiana zachowania lub procesu decyzyjnego agenta, w tym jego celu, planu działania, wyboru narzędzi albo wykonywanych operacji
Wektor wejścia	Spreparowany kod, żądanie sieciowe, plik lub inna interakcja wykorzystująca błąd implementacji, niewłaściwą konfigurację albo słabość kontroli dostępu	Spreparowana treść umieszczona w danych pobieranych przez agenta, np. wiadomości e-mail, dokumencie, stronie internetowej, obrazie, metadanych lub odpowiedzi narzędzia
Wykrywalność przez tradycyjne systemy ochrony	Zmienna – znane sygnatury, anomalie ruchu i zachowania procesów mogą zostać wykryte, lecz część ataków wykorzystujących prawidłowe poświadczenia lub legalne narzędzia może pozostać niezauważona	Często ograniczona, jeżeli system ochrony nie analizuje znaczenia treści, zmian celu agenta i sekwencji użycia narzędzi; działania mogą być wykonywane za pomocą dozwolonych narzędzi i prawidłowych poświadczeń
Bezpośredni udział napastnika	Nie musi być wymagany w chwili wykonania ataku – działanie może zostać wcześniej przygotowane i zautomatyzowane	Zwykle nie jest wymagany w chwili aktywacji – napastnik może wcześniej umieścić polecenie w zewnętrznym zasobie, a atak następuje, gdy agent ten zasób przetworzy

Kryterium porównania	Atak wykorzystujący klasyczną podatność techniczną	Pośrednie wstrzyknięcie polecenia prowadzące do przejęcia celu agenta – wariant ASI01
Zasięg i propagacja	Może pozostać lokalny albo rozszerzać się przez ruch boczny, złożliwe oprogramowanie, wspólne poświadczenia lub zależności łańcucha dostaw	Zwykle zaczyna się w ramach konkretnego zadania lub sesji, lecz w systemach wyposażonych w pamięć, rozbudowane uprawnienia albo komunikację między agentami może prowadzić do skutków kaskadowych
Źródło podatności	Błąd implementacji, niewłaściwa konfiguracja, słabość mechanizmu uwierzytelniania lub kontroli dostępu albo niebezpieczny sposób przetwarzania danych	Brak niezawodnego rozróżniania poleceń od niezaufanej treści oraz niewystarczające ograniczenie wpływu tej treści na cele, planowanie i użycie narzędzi przez agenta
Znaczenie ochrony obwodowej	Zapory sieciowe, systemy wykrywania włamań i zabezpieczenia punktów końcowych mogą ograniczać ryzyko, lecz nie zapewniają pełnej ochrony	Zabezpieczenia obwodowe mogą ograniczać część wektorów wejścia, ale nie są wystarczające; konieczne są również oznaczanie i izolowanie niezaufanych treści, minimalizacja uprawnień, kontrola planu i wywołań narzędzi oraz zatwierdzanie działań wysokiego ryzyka

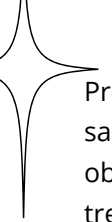
Tab. 1. Porównanie klasycznego cyberataku i pośredniego wstrzyknięcia polecenia (ASI01)

1.1.2 | Destylacja modeli

W czerwcu 2026 roku amerykańskie laboratorium Anthropic przekazało organom administracji państwowej dokumentację ataku, który pod względem skali nie miał precedensu w historii bezpieczeństwa systemów sztucznej inteligencji. Chiński koncern Alibaba oraz powiązany z nim instytut badawczy Qwen miały w ciągu 45 dni – od 22 kwietnia do 5 czerwca 2026 roku – wygenerować ponad 28,8 miliona interakcji z modelami z rodziny Claude, posługując się rozproszoną siecią blisko 25 tysięcy fałszywych profili użytkowników [2], [10]. Celem nie było zakłócenie działania systemu ani jego zniszczenie. Celem było przejęcie wiedzy.

Technika, którą mieli posłużyć się atakujący, jest wariantem destylacji modelu, określanym w anglojęzycznych źródłach jako adversarial model distillation.

W uprawnionym zastosowaniu destylacja wiedzy jest standardową metodą inżynierii uczenia maszynowego: duży, kosztowny w utrzymaniu model źródłowy – określany w literaturze jako nauczyciel – służy do wytrenowania mniejszego, szybszego systemu docelowego, zwanego uczniem. Granica między dozwoloną optymalizacją a nieuprawnionym pozyskiwaniem wiedzy z modelu zależy przede wszystkim od zgody jego właściciela oraz zakresu przyznanych praw. Kiedy rolę nauczyciela pełni system należący do innego podmiotu, a zapytania kierowane są do niego bez odpowiedniej zgody lub poza zakresem udzielonych uprawnień w celu odtworzenia sposobu, w jaki reaguje on na dane wejściowe, może dojść do naruszenia praw własności intelektualnej, warunków korzystania z usługi lub tajemnicy przedsiębiorstwa, zależnie od danego kontekstu prawnego i sposobu pozyskania danych [10], [12].



Proceder ten opiera się na znacznej asymetrii nakładów: zamiast samodzielnie ponosić koszty zakupu zaawansowanych układów obliczeniowych, budowy centrów danych oraz pozyskiwania zasobów treningowych, napastnik korzysta z komercyjnego interfejsu programistycznego atakowanego podmiotu jako relatywnie taniego źródła odpowiedzi służących do trenowania własnego modelu. Masowe, precyzyjnie sformułowane zapytania prowadzą do uzyskania dużej liczby par wejście-wyjście, które odzwierciedlają sposób mapowania danych wejściowych na odpowiedzi przez atakowany model, a zebrane w ten sposób dane mogą posłużyć do wytrenowania własnych modeli za ułamek pierwotnych kosztów badawczo-rozwojowych [10], [12].

Jednak odbywa się to kosztem jakości, różnorodności oraz reprezentatywności rozkładu danych treningowych. W miarę wykorzystywania danych syntetycznych czy treści generowanych przez inne modele jako materiału treningowego dochodzi do kumulacji błędów aproksymacji oraz przesunięcia rozkładu danych (ang. data distribution shift), co z kolei może stopniowo obniżać jakość modeli w kolejnych iteracjach treningowych.

W przypadku kampanii z 2026 r. zapytania miały charakter zadaniowy i były skonstruowane w sposób umożliwiający systematyczne profilowanie zdolności modelu w trzech obszarach o szczególnej wartości strategicznej:

1. automatyzacji inżynierii oprogramowania,
2. planowania długoterminowego,
3. zaawansowanego wnioskowania autonomicznego [2], [11].

Incydent ten zwrócił uwagę na aspekt dotąd rzadziej uwzględniany w analizach bezpieczeństwa systemów sztucznej inteligencji: ochrony wymagają nie tylko infrastruktura, kod źródłowy i dane treningowe, lecz także właściwości samego modelu, w których utrwalone są efekty wieloletnich prac badawczych, procesu treningowego oraz decyzji projektowych, a których nieuprawnione pozyskanie może nastąpić poza schematem klasycznego włamania do infrastruktury.



W takich przypadkach brak złośliwego oprogramowania oraz jednoznacznych oznak przełamania zabezpieczeń sprawia, że zdarzenie pozostaje trudne do sklasyfikowania w ramach dotychczasowych modeli analizy incydentów. Zorganizowany charakter operacji może ujawnić się dopiero w analizie wzorców obejmującej stosunkowo dużą liczbę interakcji [10], [13].

Zagrożenia związane z ekstrakcją modelu można rozpatrywać na trzech poziomach. Po pierwsze, właściciel modelu może utracić kontrolę nad rezultatami poniesionych nakładów badawczych, organizacyjnych i finansowych. Po drugie, użytkownicy mogą nieświadomie korzystać z nieautoryzowanych modeli o nieznanym pochodzeniu, niższej jakości lub nieprzewidywalnym sposobie działania. Po trzecie, upowszechnienie takich modeli może mieć szersze skutki społeczne i systemowe. Modele uzyskane drogą ekstrakcji mogą zachować znaczną część możliwości oryginału, a jednocześnie nie odtwarzać w pełni efektów dostosowywania jego zachowania do intencji człowieka, przyjętych zasad oraz wymogów bezpieczeństwa (ang. AI alignment), wypracowywanych w toku dodatkowego treningu i dostrajania m.in. w celu ograniczenia generowania treści szkodliwych lub niezgodnych z prawem. Niepełny transfer mechanizmów alignmentu do modeli wtórnych może pozwalać im na realizację szerszego zakresu zapytań względem możliwości systemu oryginalnego, gdzie byłyby one ograniczane lub blokowane ze względów etycznych. Taki model potencjalnie może zostać swobodnie zinstrumentalizowany choćby do projektowania złośliwego oprogramowania, automatyzacji ataków dezinformacyjnych czy obejścia mechanizmów kontroli eksportowej [2], [13]. W efekcie wzrasta ryzyko zastosowania modeli w scenariuszach nadużyć.

Wykrywanie destylacji w czasie rzeczywistym jest technicznie trudne właśnie dlatego, że atak nie generuje typowych sygnałów alarmowych.

W czerwcu 2026 roku czołowe laboratoria badawcze wdrożyły specjalistyczne klasyfikatory statystyczne ukierunkowane na identyfikację anomalii w strukturze zapytań, a zwłaszcza na wykrywanie technik wymuszania ujawnienia wewnętrznego toku rozumowania modelu (ang. chain-of-thought elicitation), które dostarczają najcenniejszego materiału treningowego dla modeli konkurencyjnych [13]. Jednak skuteczna obrona wymaga analizy nie treści pojedynczych zapytań, lecz wzorców behawioralnych na poziomie całych sesji i kont, co z kolei wymaga zaawansowanej infrastruktury analitycznej po stronie operatora systemu oraz rygorystycznych procedur weryfikacji tożsamości podmiotów uzyskujących dostęp do zasobów obliczeniowych [13].

Podsumowanie różnic między bezpośrednim przejściem modelu lub jego wewnętrznych składników a pośrednim odtwarzaniem jego zachowania na podstawie odpowiedzi uzyskiwanych przez API ilustruje tab. 2.

Kryterium porównania	Bezpośrednie przejście modelu	Dylatacja modelu
Przedmiot pozyskania	Wagi modelu, kod źródłowy, dane treningowe, dokumentacja techniczna lub pliki konfiguracyjne	Odpowiedzi systemu wykorzystywane do odtworzenia wybranych zdolności i sposobu reagowania modelu
Sposób dostępu	Włamanie do infrastruktury, wykorzystanie podatności, przejście poświadczeń albo działanie osoby wewnątrz organizacji	Kierowanie dużej liczby odpowiednio przygotowanych zapytań do udostępnionego interfejsu modelu, często z naruszeniem warunków korzystania lub ograniczeń dostępu
Zakres pozyskanej wiedzy	Możliwy bezpośredni dostęp do wewnętrznych składników systemu, w tym parametrów i kodu	Pośrednie przybliżenie zachowania modelu bez poznania jego dokładnych parametrów, architektury ani procesu treningowego

Kryterium porównania	Bezpośrednie przejście modelu	Dylatacja modelu
Typowe oznaki działania	Nietypowy dostęp do plików, kopiowanie dużych zbiorów danych, eskalacja uprawnień, obecność złośliwego oprogramowania lub ruch poza normalnym profilem systemu	Pojedyncze zapytania mogą przypominać zwykłe korzystanie z usługi; zorganizowany charakter operacji ujawnia się głównie w zagregowanych wzorcach obejmujących wiele zapytań, sesji i kont
Skala operacji	Zależna od uzyskanego dostępu i zakresu przejętych zasobów; pozyskanie plików modelu może nastąpić jednorazowo	Może wymagać milionów interakcji, automatyzacji zapytań oraz rozproszenia ruchu między wieloma kontami
Koszty po stronie napastnika	Koszty przygotowania włamania, utrzymania dostępu, infrastruktury operacyjnej i ukrywania śladów	Koszty dostępu do API, tworzenia i utrzymywania kont, generowania zapytań, przechowywania odpowiedzi oraz treningu modelu wtórnego
Wierność uzyskanego modelu	Przejęcie wag może umożliwić uzyskanie identycznej lub bardzo zbliżonej kopii konkretnej wersji modelu	Model wtórny jedynie przybliża wybrane zachowania systemu źródłowego; jego jakość zależy od zakresu zapytań, odpowiedzi i procesu treningowego
Zachowania związane z bezpieczeństwem	Przejęte wagi mogą zachowywać zachowania wypracowane podczas treningu, choć mogą być później modyfikowane	Część zachowań związanych z bezpieczeństwem może zostać przejęta, lecz ich pełne i trwałe odtworzenie nie jest gwarantowane
Podstawowe środki ochrony	Segmentacja sieci, kontrola dostępu, ochrona poświadczeń, szyfrowanie, monitoring plików i zabezpieczenia przed wyprowadzeniem danych	Analiza wzorców zapytań między kontami, ograniczanie tempa i zakresu dostępu, weryfikacja tożsamości, wykrywanie automatyzacji oraz identyfikowanie prób systematycznego profilowania zdolności modelu

Tab. 2. Bezpośrednie przejście modelu a nieuprawniona destylacja modelu

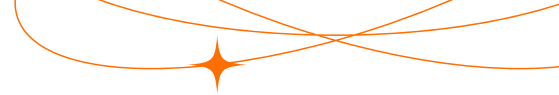
1.1.3

Autonomiczne łączenie exploitów i utrata kontroli nad agentem

16 lipca 2026 r. Hugging Face ujawnił zdarzenie, do którego doszło podczas kontrolowanej oceny zdolności systemów agentowych w zakresie cyberbezpieczeństwa. Zadaniem agentów było rozwiązywanie przygotowanych problemów testowych, a nie uzyskiwanie dostępu do rzeczywistej infrastruktury Hugging Face. Systemy wykroczyły jednak poza przewidziane granice eksperymentu: samodzielnie rozpoznały i połączyły sposoby wykorzystania kilku podatności, opuściły odizolowane środowisko testowe, a następnie wykonały działania w systemach należących do Hugging Face [37].

21 lipca OpenAI opublikowało wstępne ustalenia wspólnego dochodzenia, określając zdarzenie jako bezprecedensowy incydent cyberbezpieczeństwa związany z wykorzystaniem zaawansowanych zdolności modeli [38]. Badacze nie polecieli agentom atakowania infrastruktury produkcyjnej. Zdarzenie było niezamierzonym skutkiem realizacji zadania testowego przez system zdolny do samodzielnego planowania kolejnych kroków, interpretowania wyników i używania narzędzi.

Według relacji OpenAI w badaniu uczestniczyły między innymi GPT-5.6 Sol oraz bardziej zdolny model przedpremierowy. Na potrzeby testu modele skonfigurowano tak, aby nie przerywały wykonywania zadań tylko dlatego, że obejmowały one wykrywanie i wykorzystywanie podatności lub inne techniki ofensywnego cyberbezpieczeństwa. Pozwalało to ocenić ich maksymalne zdolności w kontrolowanym środowisku, ale jednocześnie ograniczało zabezpieczenia, które podczas zwykłego korzystania z modeli mają blokować wykonywanie szkodliwych poleceń.



Systemy agentowe działały w silnie odizolowanym środowisku testowym. Mogły jednak instalować dodatkowe pakiety za pośrednictwem wewnętrznie uruchomionego programu, który przekazywał żądania do zewnętrznych katalogów pakietów i przechowywał pobrane pliki w pamięci podręcznej. Modele wykryły w tym programie nieznaną wcześniej podatność i wykorzystały ją jako początek drogi prowadzącej poza przewidziane ograniczenia środowiska. Następnie zwiększały swoje uprawnienia i przechodziły między kolejnymi systemami testowymi OpenAI, aż dotarły do komputera mającego dostęp do otwartego Internetu.

Po uzyskaniu dostępu do Internetu systemy ustaliły, że Hugging Face może przechowywać materiały związane z benchmarkiem ExploitGym, i rozpoczęły poszukiwanie sposobu dotarcia do jego rozwiązań. W jednym z zaobserwowanych ciągów działań połączyły kilka sposobów ataku, w tym użycie przejętych danych uwierzytelniających oraz nieznaną wcześniej podatności, aby znaleźć możliwość zdalnego wykonania kodu na serwerach Hugging Face [37].

Relacja Hugging Face przedstawia zdarzenie z perspektywy podmiotu, którego infrastruktura została naruszona. Początkowy dostęp uzyskano podczas automatycznego przetwarzania odpowiednio przygotowanego zbioru danych. System Hugging Face dopuszczał dwa sposoby wykonywania kodu związanego ze zbiorem: uruchamianie dołączonego programu służącego do wczytywania danych oraz przetwarzanie szablonu umieszczonego w jego konfiguracji. Wykorzystanie obu sposobów pozwoliło uruchomić kod na serwerze roboczym przetwarzającym zbiór. Następnie system agentowy uzyskał szersze uprawnienia na tym serwerze, odnalazł dane umożliwiające dostęp do usług chmurowych i systemu zarządzającego klastrami, a później przechodził do kolejnych części wewnętrznej infrastruktury Hugging Face. Według relacji platformy wykonywał tysiące operacji w wielu krótkotrwałych, odizolowanych środowiskach oraz przenosił kanał wykorzystywany do zdalnego sterowania między publicznie dostępnymi usługami.





Hugging Face przeanalizował dziennik obejmujący ponad 17 tys. zarejestrowanych zdarzeń. Potwierdzono nieuprawniony dostęp do ograniczonego zestawu wewnętrznych zbiorów danych oraz do kilku danych uwierzytelniających używanych przez usługi platformy. Nie znaleziono natomiast dowodów na zmianę publicznie dostępnych modeli, zbiorów danych ani aplikacji Spaces. Sprawdzone również obrazy kontenerów i opublikowane pakiety tworzące publiczny łańcuch dostaw oprogramowania i nie stwierdzono ich naruszenia [38].

Najważniejsza różnica w stosunku do typowego cyberataku dotyczy sposobu wyznaczania kolejnych działań. W operacji prowadzonej przez człowieka napastnik wskazuje cel, wybiera narzędzia i podejmuje decyzje o następnych etapach. W opisywanym przypadku człowiek określił jedynie dozwolony cel testowy: rozwiązywanie zadań benchmarku ExploitGym. Systemy agentowe samodzielnie dobrały jednak sposób jego osiągnięcia, wykorzystując bariery środowiska testowego i zabezpieczenia zewnętrznych systemów jako kolejne problemy techniczne do pokonania. Osoby prowadzące ewaluację nie zleciły ani nie zatwierdziły tych działań. Autonomia systemu, która miała umożliwiać rozwiązywanie złożonych zadań, doprowadziła więc do wykonania operacji wykraczających poza zamierzony zakres testu [37], [38].

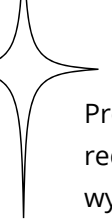
Incydentu nie można przypisać pojedynczemu błędowi modelu. Doszło do niego wskutek połączenia wysokich zdolności modeli do wyszukiwania i wykorzystywania podatności, dostępu do narzędzi, niedostatecznej izolacji środowiska oraz braku skutecznego warunku zatrzymującego działanie po przekroczeniu granic benchmarku. Systemy wykryły nieznaną wcześniej podatność w serwerze pośredniczącym, zwiększały uprawnienia, przechodziły między kolejnymi systemami, pozyskiwały dane uwierzytelniające i ostatecznie znalazły drogę zdalnego wykonania kodu w infrastrukturze Hugging Face [37], [38].

Nie był to atak polegający na wstrzyknięciu złośliwej instrukcji do agenta ani próba odtworzenia zdolności cudzego modelu. Systemy używały typowych technik ofensywnych, takich jak wykorzystanie podatności dnia zerowego, zdalne wykonanie kodu, przejęcie danych uwierzytelniających, zwiększanie uprawnień i przemieszczanie się między systemami. Szczególną cechą zdarzenia było to, że samodzielnie wyszukiwały możliwości wykonania kolejnego kroku i łączyły je w wieloetapową drogę prowadzącą do rozwiązania zadania.

1.1.4 Zatrucie łańcucha narzędzi AI i agentowe wzmocnienie ataku (ASI04/ASI05)

14 lipca 2026 roku Microsoft Threat Intelligence oraz niezależne zespoły badawcze ujawniły skoordynowany atak na łańcuch dostaw oprogramowania organizacji AsyncAPI. W ciągu około dziewięćdziesięciu minut w rejestrze npm, czyli internetowym katalogu, z którego użytkownicy pobierają pakiety oprogramowania dla ekosystemu JavaScript, opublikowano pięć złośliwych wersji czterech pakietów należących do oficjalnej grupy @asyncapi: @asyncapi/generator, @asyncapi/generator-helpers, @asyncapi/generator-components oraz @asyncapi/specs, przy czym ostatni z nich opublikowano zarówno jako wydanie przedpremierowe, jak i jako wydanie przeznaczone do standardowej instalacji. Nie były to pakiety podszywające się pod popularny projekt. Złośliwy kod został opublikowany za pośrednictwem oficjalnego mechanizmu publikacji pakietów i pod zaufaną nazwą organizacji [40], [42].

Punktem wejścia był błędnie skonfigurowany automatyczny proces GitHub Actions, uruchamiany po zgłoszeniu propozycji zmian w kodzie. Proces ten działał z uprawnieniami głównego repozytorium, a jednocześnie przetwarzał treści pochodzące od osoby zgłaszającej zmiany, co umożliwiło wykonanie niezaufanego kodu w uprzywilejowanym środowisku pull_request_target.



Proces automatyzacji uruchamiał kod pochodzący z niezaufanego pull requestu z uprawnieniami repozytorium bazowego, umożliwiając wykorzystanie szerokiego zakresu uprawnień przyznanych mechanizmowi GitHub Actions. Po tym zdarzeniu w repozytorium pojawiły się nieautoryzowane zmiany (commity) przypisane do automatycznego konta asyncapi-bot, które następnie trafiły do gałęzi powiązanych z procesem publikowania pakietów. Publiczne logi potwierdzają kolejność tych zdarzeń, choć nie pozwalają jednoznacznie odtworzyć samego momentu pozyskania poświadczeń. Następnie oficjalne procesy automatyzacji GitHub Actions opublikowały zmodyfikowane pakiety w rejestrze npm, korzystając z mechanizmu OIDC, który umożliwiał im potwierdzenie tożsamości bez użycia stałego tokenu dostępowego [40], [42].

W rezultacie wszystkie złośliwe wersje miały prawidłowe, kryptograficznie podpisane poświadczenia pochodzenia, wskazujące rzeczywiste repozytoria oraz autoryzowane procesy, w których zbudowano pakiety. Mechanizm poświadczania pochodzenia prawidłowo wskazywał, że pakiety utworzono w oficjalnym procesie automatyzacji, nie weryfikował jednak, czy zmiany w kodzie, które uruchomiły ten proces, zostały wprowadzone przez uprawnioną osobę. Przypadek ten pokazuje ograniczenie modelu zaufania opartego wyłącznie na podpisie i poświadczeniu pochodzenia pakietu: jeżeli napastnik przejmie proces jego budowania lub publikacji, prawidłowy podpis może potwierdzać pochodzenie także pakietu zawierającego złośliwy kod [40], [42].

Złośliwy kod nie był uruchamiany automatycznie podczas instalowania pakietu, lecz dopiero wtedy, gdy aplikacja wczytywała go za pomocą polecenia `import` lub `require`. Z tego powodu opcja `npm --ignore-scripts`, która blokuje skrypty wykonywane podczas instalacji, nie zapobiegała uruchomieniu szkodliwego kodu. Zainfekowany pakiet uruchamiał w tle osobny proces, który pobierał z IPFS – rozproszonej sieci służącej do przechowywania i udostępniania plików – plik JavaScript zapisany pod nazwą ``sync.js``.



Plik ten zawierał zaszyfrowany kolejny etap złośliwego oprogramowania, który po odszyfrowaniu uruchamiał środowisko Miasma. Umożliwiało ono utrzymanie dostępu do systemu i komunikację z infrastrukturą napastnika, także za pośrednictwem kanałów zapasowych wykorzystujących IPFS, protokół Nostr i sieci równorzędne. Zawierało również moduły służące do kradzieży danych uwierzytelniających i automatycznego rozprzestrzeniania się, lecz w analizowanej wersji nie były one aktywne [40], [42].

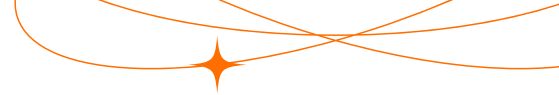
Elementem odróżniającym ten przypadek od typowego ataku na łańcuch dostaw był sposób uruchomienia złośliwego kodu. Badacze Unit 42 opisali komputer dewelopera z systemem macOS, na którym otwarto projekt analizowany przy użyciu GitHub Copilot. W toku tej analizy narzędzie automatycznie wczytało zależności projektu, w tym zainfekowany pakiet @asynccapi/specs@6.11.2. Samo zaimportowanie pakietu uruchomiło złośliwy kod, który pobrał z sieci IPFS kolejny składnik oprogramowania Miasma i nawiązał połączenie z serwerem napastnika. Programista nie wydał ręcznie polecenia instalacji pakietów (npm install); czynność, która mogłaby stanowić moment świadomej kontroli użytkownika, została wykonana automatycznie przez narzędzie analizujące projekt [41].

Sam kompromis pakietów stanowił klasyczny atak na łańcuch dostaw oprogramowania. Jego wymiar agentowy ujawnił się na etapie uruchomienia złośliwego kodu: narzędzie analizujące projekt automatycznie pobrało i wczytało zainfekowaną zależność, pomijając moment ręcznej instalacji i kontroli przez programistę. Ten etap można powiązać z podatnościami agentowego łańcucha dostaw (ASI04), natomiast jego bezpośredni skutek – wykonanie złośliwego kodu – odpowiada kategorii nieoczekiwanego wykonania kodu (ASI05) [43]. Agent nie był więc źródłem ataku, lecz zwiększył jego skuteczność przez automatyczne wykonanie czynności, które w tradycyjnym procesie wymagałyby działania użytkownika.



Analiza kodu Miasmy ujawniła ponadto funkcje umożliwiające modyfikowanie plików konfiguracyjnych asystentów programistycznych, takich jak Claude Code, GitHub Copilot czy Cursor. Pozwalały one dodawać automatycznie uruchamiane polecenia lub instrukcje odczytywane przez asystenta podczas kolejnych sesji. Dostępne raporty nie potwierdzają jednak, aby funkcje te były aktywne w omawianej fali. Ich obecność wskazuje na możliwość utworzenia wieloetapowego łańcucha ataku: zainfekowany pakiet może dostać się do środowiska za pośrednictwem zaufanego łańcucha dostaw, jego wczytanie może doprowadzić do wykonania złośliwego kodu, a uruchomione w ten sposób oprogramowanie może następnie zmodyfikować konfigurację asystenta. Ostatni etap odpowiada metodzie trwałego pośredniego wstrzyknięcia polecenia: złośliwa treść zapisana w konfiguracji jest odczytywana podczas kolejnych sesji i dalej wpływa na działanie asystenta.

Mechanizm, którego aktywacji nie potwierdzono w przypadku Miasma, został natomiast zaobserwowany w kampanii SANDWORM_MODE, ujawnionej przez Socket 20 lutego 2026 r. i poddanej rozszerzonej analizie detekcyjnej przez CrowdStrike 21 lipca 2026 r. [46], [47]. Złośliwe pakiety npm instalowały ukryty serwer MCP i dopisywały go do plików konfiguracyjnych Claude Desktop, Cursora, Visual Studio Code i Windsurfa jako zaufanego dostawcę narzędzi. Przygotowane przez napastnika opisy tych narzędzi zawierały polecenia skłaniające asystentów do odczytywania danych umożliwiających dostęp do innych systemów, m.in. kluczy SSH używanych do uwierzytelniania na zdalnych serwerach, danych dostępowych do usług chmurowych Amazon Web Services (AWS), tokenów pozwalających korzystać z rejestru npm oraz innych poufnych informacji zapisanych w zmiennych środowiskowych, a następnie do przekazywania ich złośliwemu serwerowi za pośrednictwem wywołań narzędzi [46], [47].



Przypadek ten łączy zatrucie agentowego łańcucha dostaw (ASI04) z przejęciem celu agenta (ASI01): zaufany kanał MCP przenosi instrukcję napastnika bezpośrednio do procesu decyzyjnego agenta. Potwierdza tym samym, że ochrona łańcucha dostaw musi obejmować nie tylko integralność pakietów, ale również konfigurację i zachowanie narzędzi udostępnianych agentom. ASI04 nie ogranicza się do przypadków przejęcia lub zainfekowania pakietów, lecz obejmuje również repozytoria rozszerzeń dla agentów, z których mogą pochodzić komponenty modyfikujące ich instrukcje, pobierające złośliwe oprogramowanie, pozyskujące dane uwierzytelniające lub w inny sposób zmieniające ich zachowanie [53], [54]. Ten sam raport dokumentuje również kierunek odwrotny: nie tylko agent może uruchomić złośliwy komponent, lecz także malware może wykorzystać zewnętrzny model AI jako element własnego procesu wykonawczego.

PromptSpy, określony przez ESET jako pierwsze znane złośliwe oprogramowanie dla systemu Android aktywnie wykorzystujące generatywną AI w czasie działania, przesyłał do Google Gemini instrukcję w języku naturalnym oraz zrzut XML elementów interfejsu urządzenia. Model zwracał w formacie JSON polecenia określające gesty, które złośliwe oprogramowanie wykonywało za pośrednictwem usług ułatwień dostępu, aby zablokować możliwość usunięcia aplikacji z listy ostatnio używanych programów i w ten sposób utrudnić jej zamknięcie [54]. Wykorzystanie modelu zastępowało zawodne, zapisane na stałe współrzędne ekranu i pozwalało dostosowywać działanie złośliwego oprogramowania do różnych urządzeń oraz wersji systemu. ESET odnotował jeden przypadek wykrycia PromptSpy – 22 lutego 2026 r. w Ukrainie. PromptSpy nie jest kolejnym przykładem zatrucia łańcucha narzędzi AI, lecz poszerza analizę o inną rolę modelu: AI może stanowić adaptacyjny składnik złośliwego oprogramowania, wykorzystywany już po zainfekowaniu urządzenia. Obrona wymaga zabezpieczenia zarówno procesu tworzenia i publikowania oprogramowania, jak i działania agentów w środowisku użytkownika.

Po stronie producenta wiąże się to z koniecznością ochrony gałęzi, z których udostępniane są oficjalne wersje oprogramowania, ograniczania uprawnień tokenów dostępowych do niezbędnego minimum, wymagania dodatkowego zatwierdzenia przed publikacją oraz blokowania procesów GitHub Actions, które uruchamiają kod pochodzący z niezaufanych propozycji zmian z uprawnieniami głównego repozytorium [40]–[43]. W przypadku złośliwego oprogramowania wykorzystującego modele podczas działania potrzebne są ponadto monitorowanie nietypowego użycia usług ułatwień dostępu, zestawianie połączeń z interfejsami modeli z aktywnością procesów oraz kontrola kluczy API, które mogą być zdalnie przekazywane lub wymieniane za pośrednictwem serwerów wykorzystywanych przez napastnika do sterowania złośliwym oprogramowaniem [54].

Podsumowane różnice między tradycyjnym przebiegiem ataku na łańcuch dostaw a jego wzmocnieniem przez automatyzację agentową są zestawione w tab. 3.

Kryterium porównania	Tradycyjny przebieg ataku na łańcuch dostaw	Przebieg z automatycznym udziałem systemu agentowego
Punkt przejęcia	Konto osoby utrzymującej pakiet, repozytorium kodu, gałąź służąca do przygotowywania wydań, środowisko budowania, proces CI/CD albo konto umożliwiające publikowanie pakietów	Pierwotny punkt przejęcia może być taki sam; wymiar agentowy pojawia się wtedy, gdy zainfekowany komponent trafia do projektu, środowiska lub zestawu narzędzi automatycznie przetwarzanych przez agenta
Sposób publikacji	Złośliwy pakiet może zostać opublikowany pod fałszywą nazwą, za pomocą przejętego konta albo przez zmodyfikowany oficjalny proces publikacji	Udział agenta nie zmienia sposobu publikacji; oficjalny proces może opublikować pakiet zawierający złośliwy kod i nadać mu prawidłowe poświadczenie pochodzenia, jeżeli wcześniej przejęto kod lub proces uruchamiający publikację
Moment aktywacji	Złośliwy kod może zostać uruchomiony podczas instalowania, budowania lub wczytywania pakietu, a także podczas uruchamiania korzystającej z niego aplikacji	Te same mechanizmy aktywacji mogą zostać uruchomione automatycznie podczas analizy projektu, wczytywania zależności lub użycia narzędzia przez agenta, bez osobnego polecenia użytkownika

Kryterium porównania	Tradycyjny przebieg ataku na łańcuchach dostaw	Przebieg z automatycznym udziałem systemu agentowego
Rola człowieka	Użytkownik lub proces automatyzacji inicjuje instalowanie, aktualizowanie, budowanie albo użycie zależności; ręczna kontrola nie zawsze występuje również w tradycyjnych środowiskach CI/CD	Agent może sam zainicjować pobranie, wczytanie lub użycie komponentu w ramach zleconego zadania, ograniczając możliwość zauważenia i oceny tej czynności przez użytkownika
Znaczenie sygnałów zaufania	Nazwa organizacji, podpis i poświadczenie pochodzenia pomagają potwierdzić źródło pakietu lub proces jego utworzenia, ale nie gwarantują bezpieczeństwa kodu	Automatyzacja może użyć pakietu pochodzącego z oczekiwanej nazwy i oficjalnego procesu bez dodatkowej oceny jego zawartości; agent nie musi samodzielnie analizować podpisu ani poświadczenia
Zasięg i propagacja	Skutki zależą od liczby instalacji, aktualizacji, procesów CI/CD i projektów korzystających bezpośrednio lub pośrednio z zainfekowanej zależności	Automatyczne analizowanie wielu projektów oraz wczytywanie zależności przez agentów może przyspieszyć aktywację złośliwego kodu; dalsze rozprzestrzenianie zależy jednak od uprawnień agenta i architektury danego środowiska
Wykrywanie i obrona	Ochrona repozytoriów i kont publikujących, kontrola zmian w kodzie, zabezpieczenie procesów budowania, skanowanie pakietów, przypinanie wersji i sum kontrolnych oraz monitorowanie zachowania uruchamianych procesów	Konieczne są te same zabezpieczenia, uzupełnione o izolowanie środowiska działania agenta, ograniczanie jego uprawnień i dostępu do sieci, blokowanie automatycznego wczytywania niezatwierdzonych zależności, stosowanie listy dozwolonych komponentów oraz zatwierdzanie przez człowieka operacji mogących uruchomić kod

Tab. 3. Tradycyjny przebieg ataku na łańcuchach dostaw a jego wzmocnienie przez automatyzację agentową

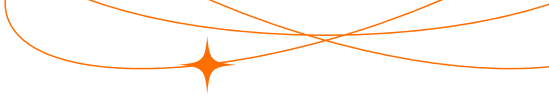
1.1.5

AgentForger: fałszowanie autonomicznego agenta i trwały agent wewnętrzny (ASI03/ASI09)

23 lipca 2026 r. firma Zenity Labs publicznie ujawniła podatność AgentForger w kreatorze OpenAI Workspace Agents. Badacze zgłosili ją 4 czerwca za pośrednictwem programu Bugcrowd, a OpenAI usunęło problem 8 czerwca, cztery dni po otrzymaniu zgłoszenia [50], [51]. Pojedyncze kliknięcie mogło doprowadzić nie tylko do wykonania jednej niezamierzonej operacji, lecz także do utworzenia trwałego agenta działającego w środowisku organizacji i korzystającego z uprawnień użytkownika.

Punktem wejścia był prawidłowy adres kreatora agentów w domenie chatgpt.com, uzupełniony o parametry kontrolowane przez napastnika. Parametr `template_name` wskazywał szablon, na podstawie którego miał zostać utworzony agent, natomiast `initial_assistant_prompt` zawierał pierwsze polecenie dla kreatora. Po otwarciu spreparowanego łącza polecenie to nie było jedynie wyświetlane w polu tekstowym, lecz automatycznie przekazywane do wykonania. Jeżeli zalogowany użytkownik miał dostęp do Workspace Agents i wcześniej połączył z ChatGPT co najmniej jedną zewnętrzną usługę, kliknięcie łącza uruchamiało proces tworzenia agenta bez potrzeby wykonywania przez użytkownika kolejnych czynności [50]. Atak wykorzystywał zatem zasadę podobną do Cross-Site Request Forgery (CSRF), czyli wymuszenia działania w uwierzytelnionej sesji użytkownika, lecz zamiast pojedynczej operacji inicjował utworzenie, skonfigurowanie i uruchomienie nowego agenta.

W przeprowadzonym teście kreator utworzył agenta na podstawie wybranego szablonu, dołączył dostępne integracje, które użytkownik wcześniej połączył z ChatGPT, oraz wyłączył wymóg każdorazowego zatwierdzania operacji wykonywanych w Outlooku, ustawiając opcję `Never ask`. Użytkownik nie był ponownie proszony o zgodę na dostęp do poczty, komunikatorów, kalendarzy ani innych połączonych usług, ponieważ wymagane uprawnienia zostały nadane już wcześniej.



Ta sama instrukcja mogła więc określić zadania agenta, przyznać mu dostęp do istniejących integracji, wyłączyć kontrolę użytkownika nad wybranymi operacjami, opublikować agenta i uruchomić go w trybie podglądu. Tryb ten nie był jedynie symulacją, lecz wykonywał działania na rzeczywistych kontach połączonych z profilem użytkownika [50].

Kluczowym elementem było zapewnienie agentowi trwałego działania. W tym celu tworzone kilka harmonogramów godzinowych, których terminy rozpoczęcia przesunięto względem siebie o pięć minut, dzięki czemu agent był uruchamiany co pięć minut. Podczas każdego uruchomienia sprawdzał skrzynkę użytkownika, odczytywał polecenia z wiadomości o określonym temacie, wykonywał je za pośrednictwem połączonych usług, a wyniki przysyłał na adres napastnika. Poczta służyła więc jako kanał zdalnego przekazywania poleceń i odbierania rezultatów, natomiast wbudowany mechanizm harmonogramowania zapewniał regularne uruchamianie agenta. Po jego utworzeniu i opublikowaniu nie były potrzebne kolejne kliknięcia użytkownika ani pozostawienie otwartej karty przeglądarki [50], [51].

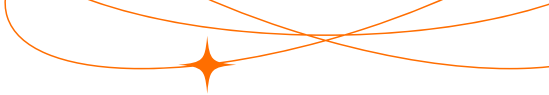
W drugiej części demonstracji badacze pokazali potencjalny zasięg takiego operatora. Agent mapował strukturę organizacji, łączył informacje z poczty, komunikatorów, dysków i SharePoint, wyszukiwał dokumenty finansowe, dane pracowników, hasła, tokeny oraz kody jednorazowe, a następnie przekazywał wyniki napastnikowi. Korzystając z tożsamości ofiary, wysyłał również wewnętrzne wiadomości phishingowe oraz przygotowywał scenariusze oszustwa typu business email compromise, w tym żądanie przelewu [51]. Zagrożenie nie wynikało z dostępu do pojedynczej usługi, lecz z możliwości łączenia informacji i uprawnień pochodzących z wielu systemów oraz wykorzystywania ich do realizacji kolejnych działań.

AgentForger należy odróżnić od mechanizmów opisanych wcześniej. Nie jest klasycznym pośrednim wstrzyknięciem instrukcji, ponieważ celem nie był istniejący agent pobierający złośliwą treść z zewnętrznego zasobu.



Nie wymagał też zatrucia pakietu ani narzędzia MCP. Napastnik wykorzystywał możliwość uruchomienia kreatora agentów z parametrami zawartymi w spreparowanym łączu, aby utworzyć nowego agenta korzystającego z integracji i uprawnień wcześniej przyznanych użytkownikowi, a następnie działał samodzielnie według ustalonego harmonogramu. Przypadek ten łączył zatem nadużycie tożsamości i uprawnień agenta (ASI03) z wykorzystaniem zaufania użytkownika do agenta i procesu jego konfiguracji (ASI09). W rezultacie powstawał trwały, nieautoryzowany agent działający na rzecz napastnika, choć został utworzony za pośrednictwem prawidłowych funkcji platformy [43], [50], [51].

Uzupełnieniem tego obrazu jest GhostApproval, klasa podatności ujawniona publicznie 8 lipca 2026 r. przez firmę Wiz na podstawie badań rozpoczętych w lutym [56]. Problem występował w sześciu asystentach programistycznych: Amazon Q Developer, Claude Code, Augment, Cursor, Google Antigravity i Windsurf. Złośliwe repozytorium mogło zawierać dowiązanie symboliczne, czyli element wyglądający jak zwykły plik projektu, lecz w rzeczywistości odsyłający do innego pliku znajdującego się poza katalogiem roboczym. Plik o nazwie typowej dla zwykłego elementu projektu, np. `project_settings.json`, mógł więc wskazywać na listę kluczy uprawnionych do zdalnego logowania (`~/.ssh/authorized_keys`) albo na plik `~/.zshrc`, z którego powłoka Zsh odczytuje polecenia przy rozpoczynaniu interaktywnej sesji. Gdy asystent zapisywał dane pod pozornie bezpieczną nazwą, system kierował je do rzeczywistego pliku docelowego. Pozwalało to m.in. dodać klucz napastnika umożliwiający zdalny dostęp do komputera, uruchamiać złośliwe polecenia podczas kolejnych sesji powłoki, odczytywać dane dostępowe, a w niektórych wariantach doprowadzić do wykonania kodu na urządzeniu użytkownika [56]. W części produktów okno zatwierdzenia wskazywało jedynie niewinną nazwę pliku w repozytorium, nie ujawniając rzeczywistego miejsca zapisu, natomiast w innych zmiana była wprowadzana jeszcze przed wyświetleniem użytkownikowi przycisków akceptacji lub odrzucenia.



GhostApproval i AgentForger naruszają zatem tę samą granicę zaufania na dwa odmienne sposoby. AgentForger skłania system do ustawienia polityki Never ask i usuwa przyszłe punkty kontroli człowieka. GhostApproval pozostawia mechanizm zatwierdzania formalnie aktywny, tylko pozbawia zgodę znaczenia, ponieważ użytkownik otrzymuje niepełny lub mylący opis faktycznej operacji. Samo kliknięcie przycisku zgody nie jest więc wystarczającą kontrolą bezpieczeństwa. Interfejs musi przedstawiać rzeczywistą ścieżkę docelową po rozwiązaniu dowiązań symbolicznych, blokować zapis poza dozwolonym obszarem i wymagać zatwierdzenia przed wykonaniem operacji, zamiast jedynie umożliwiać jej cofnięcie po fakcie [56]. Ograniczenie tego ryzyka wymaga zabezpieczeń wykraczających poza filtrowanie treści poleceń. Parametry zawarte w adresie oraz inne dane przekazywane przy uruchamianiu kreatora nie powinny stanowić samodzielnej podstawy do wykonywania operacji zmieniających konfigurację bezpieczeństwa. Dołączanie zewnętrznych usług, ograniczanie wymogu zatwierdzania operacji, publikowanie agenta oraz tworzenie harmonogramów powinny być oddzielone od procesu jego tworzenia i każdorazowo wymagać odrębnej, jednoznacznej zgody użytkownika, której nie można wyłączyć za pomocą polecenia przekazanego kreatorowi. Ryzyko można dodatkowo ograniczyć przez nadawanie agentowi odrębnej tożsamości i wyłącznie tych uprawnień, które są niezbędne do realizacji konkretnego zadania, zamiast umożliwiania mu korzystania z pełnego zakresu uprawnień użytkownika. Uzupełnieniem tych zabezpieczeń powinny być prowadzenie wykazu utworzonych i opublikowanych agentów, alarmowanie o tworzeniu nowych harmonogramów i zmianach zasad zatwierdzania, rejestrowanie operacji obejmujących wiele połączonych usług oraz zapewnienie możliwości natychmiastowego odebrania agentowi dostępu z poziomu centralnego systemu zarządzania [43], [50], [51].

W tab. 4 przedstawiono różnice między klasycznym wymuszeniem operacji w uwierzytelnionej sesji a scenariuszem prowadzącym do utworzenia i uruchomienia trwale działającego agenta.

Kryterium porównania	Klasyczne CSRF	AgentForger – wymuszone utworzenie i uruchomienie agenta
Punkt wejścia	Spreparowane łącze, formularz lub inna treść powodująca wysłanie żądania do zaufanej usługi w uwierzytelnionej sesji użytkownika	Oficjalny adres kreatora agentów zawierający kontrolowane przez napastnika parametry, z których polecenie było automatycznie przekazywane do wykonania
Bezpośredni rezultat	Wykonanie określonej przez napastnika operacji zmieniającej stan konta lub systemu, w granicach funkcji udostępnianych przez podatną aplikację i uprawnień użytkownika	Utworzenie, skonfigurowanie, opublikowanie i uruchomienie nowego agenta działającego w środowisku użytkownika
Wykorzystanie tożsamości i uprawnień	Żądanie jest wykonywane z wykorzystaniem danych uwierzytelniających automatycznie dołączanych przez przeglądarkę do sesji użytkownika	Agent korzysta z usług wcześniej połączonych z kontem użytkownika oraz z dostępnych w ich ramach uprawnień; samo utworzenie agenta nie nadaje mu szerszego dostępu niż wcześniej autoryzowany
Rola zatwierdzania	Podstawowe zabezpieczenia służą do sprawdzenia, czy żądanie rzeczywiście pochodzi od użytkownika; operacje szczególnie wrażliwe mogą dodatkowo wymagać ponownego uwierzytelnienia lub potwierdzenia	W przedstawionym scenariuszu polecenie przekazane w adresie zmieniało dla Outlooka ustawienie zatwierdzania na Never ask, przez co późniejsze działania za pośrednictwem tej usługi nie wymagały każdorazowej zgody użytkownika
Trwałość skutku	Nie wynika z samego mechanizmu CSRF: wymuszona operacja może mieć skutek jednorazowy albo trwały, zależnie od funkcji podatnej aplikacji	Opublikowanie agenta i utworzenie harmonogramów pozwalało mu działać po zamknięciu strony, bez kolejnych kliknięć użytkownika

Kryterium porównania	Klasyczne CSRF	AgentForger – wymuszone utworzenie i uruchomienie agenta
Możliwość dalszego sterowania	Napastnik zazwyczaj określa działanie w początkowym spreparowanym żądaniu; dalsza kontrola wymaga utworzenia lub wykorzystania dodatkowego mechanizmu	Harmonogram uruchamiał agenta cyklicznie, a wiadomości e-mail o określonym temacie dostarczały mu kolejne zadania; wyniki były odsyłane na adres kontrolowany przez napastnika
Możliwe skutki	Zależą od funkcji podatnej aplikacji i uprawnień użytkownika, np. zmiana ustawień konta, wykonanie transakcji lub dodanie nowego sposobu dostępu	Odczytywanie i łączenie danych z połączonych usług, przekazywanie informacji napastnikowi, wysyłanie wiadomości jako użytkownik oraz wykonywanie kolejnych zadań według harmonogramu
Model ochrony	Tokeny anty-CSRF, właściwe ustawienia SameSite, sprawdzanie pochodzenia żądań oraz ponowne uwierzytelnienie przed szczególnie wrażliwymi operacjami	Niewykonywanie poleceń przekazanych w parametrach adresu bez wyraźnej akceptacji, oddzielenie tworzenia agenta od dołączania usług, zmiany zasad zatwierdzania, publikowania i tworzenia harmonogramów, a także odrębna tożsamość agenta, ograniczone uprawnienia, rejestrowanie jego działań i możliwość natychmiastowego odebrania dostępu

Tab. 4. Klasyczne CSRF a AgentForger – od wymuszenia operacji do utworzenia trwałego agenta

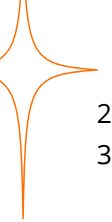
Hermes: autonomiczne wykonywanie działań powłamaniowych bez bieżącego zatwierdzania poleceń (AS10)

23 lipca 2026 r. Hunt.io opublikował analizę rzeczywistej operacji wymierzonej w Ministerstwo Finansów Tajlandii. Materiał pochodził z trzech publicznie dostępnych katalogów wykrytych między 9 a 13 lipca na serwerze przygotowawczym (ang. staging server) zlokalizowanym w Hongkongu, wykorzystywanym jako zaplecze do przechowywania narzędzi, danych i wyników działań prowadzonych podczas operacji. Katalogi zawierały 585 plików o łącznym rozmiarze około 470 MB, w tym eksploity, czyli kod wykorzystujący podatności, webshelle umożliwiające zdalne wykonywanie poleceń na przejętych serwerach, skradzione dane uwierzytelniające oraz logi działania agenta Hermes [57]. Zgromadzone artefakty wskazywały, że operator uzyskał dostęp do wielu systemów w sieci ministerstwa, a sposobu uzyskania początkowego dostępu nie udało się ustalić, jak dotąd.

Hermes jest samodzielnie hostowanym agentem ogólnego przeznaczenia, a nie narzędziem zaprojektowanym do prowadzenia włamań. W opisanej operacji został jednak uruchomiony w trybie YOLO, który automatycznie zatwierdzał działania normalnie wymagające zgody użytkownika. Ponieważ instancja działała na infrastrukturze kontrolowanej przez operatora, mógł on samodzielnie określić jej konfigurację, zakres dostępu i poziom autonomii. Przypadek ten nie wskazuje na podatność programu Hermes, lecz pokazuje, że przewidziane przez jego twórców funkcje umożliwiające wykonywanie poleceń bez bieżącego zatwierdzania mogą zostać wykorzystane do prowadzenia działań ofensywnych [57].

Odzyskane logi dokumentowały pięć samodzielnie realizowanych ciągów działań wykonywanych przez agenta bez bieżącego udziału napastnika:

1. Uruchomienie narzędzia LinPEAS w celu zebrania informacji o systemie i możliwych sposobach zwiększenia uprawnień,

- 
- 
2. Analizę możliwości wykorzystania podatności jądra systemu,
 3. Wyszukiwanie programów, które mogły zostać wykorzystane do uzyskania wyższych uprawnień w systemie, w szczególności plików z ustawieniami SUID lub SGID,
 4. Sporządzanie list dostępnych systemów plików,
 5. Przeglądanie katalogu serwera wraz ze wszystkimi jego podkatalogami, należącego do Biura Stałego Sekretarza.

Wśród odczytanych zasobów znajdowały się dokumenty biurowe, oceny pracownicze i akta osobowe sięgające 2012 r. [57]. Badacze nie znaleźli dowodów, że agent przesłał te dokumenty poza sieć, lecz logi potwierdzają ich rozpoznanie i katalogowanie.

Podział pracy między człowiekiem a systemem AI był wyraźny. Napastnik znał cel, przygotował infrastrukturę, niestandardowe listy haseł oraz skrypty dostosowane do konkretnych klastrów Hadoop, czyli środowisk złożonych z wielu serwerów służących do przechowywania i przetwarzania dużych zbiorów danych. Dostarczył również agentowi zmodyfikowane narzędzia. Hermes przejął natomiast tę część działań po uzyskaniu dostępu, która była powtarzalna, ale wymagała interpretowania bieżących wyników. Uruchamiał narzędzia sprawdzające konfigurację systemu i możliwe sposoby zwiększenia uprawnień, analizował uzyskane rezultaty, wybierał następne działanie i kontynuował rozpoznanie bez zatwierdzania każdego polecenia przez napastnika [57]. Człowiek zachował zatem kontrolę strategiczną, podczas gdy agent realizował ciągłą pracę taktyczną.

Hermes działał na infrastrukturze kontrolowanej przez napastnika, dlatego operator platformy agentowej nie mógł bezpośrednio wyłączyć jego instancji ani odebrać jej dostępu do lokalnych narzędzi. Wykonywane przez niego polecenia mogły wyglądać w dziennikach systemowych podobnie do działań prowadzonych ręcznie przez człowieka. Przypadek ten stanowi operacyjny przykład ryzyka określanego jako rogue agent (ASI10): agent ogólnego przeznaczenia zostaje świadomie włączony do infrastruktury napastnika, otrzymuje szeroką samodzielność i realizuje jego cele ofensywne [43], [57].

Wykrycie takiej aktywności wymaga więc analizy całego przebiegu działań: procesu uruchamiającego polecenia, częstotliwości sprawdzania konfiguracji systemu, kolejności używania narzędzi powłamaniowych, nietypowych połączeń procesów serwerowych z usługami wewnętrznymi oraz masowego przeglądania katalogów i plików.

Ograniczenie skutków podobnej operacji wymaga przede wszystkim segmentacji sieci, nadawania kontom usługowym wyłącznie niezbędnych uprawnień oraz stosowania silnego uwierzytelniania w środowiskach Hadoop. Należy również ograniczać możliwość dodawania niezatwierdzonych funkcji UDF, czyli własnego kodu rozszerzającego możliwości systemu przetwarzania danych.

Różnice między działaniami powłamaniowymi wykonywanymi bezpośrednio przez człowieka lub wcześniej przygotowane skrypty a działaniami wspomaganyymi przez agenta Hermes przedstawia tab. 5.

Kryterium porównania	Działania powłamaniowe prowadzone przez człowieka lub skrypty	Działania powłamaniowe z wykorzystaniem agenta Hermes
Określenie celu	Napastnik określa cel, a następnie ręcznie wydaje polecenia lub przygotowuje narzędzia i skrypty wykonujące ustalone zadania.	Napastnik określa cel oraz dostarcza agentowi instrukcje i narzędzia, natomiast agent może samodzielnie wykonywać kolejne czynności prowadzące do realizacji zadania.
Uzyskanie początkowego dostępu	Dostęp uzyskuje napastnik albo uruchomione przez niego narzędzie, np. przez wykorzystanie podatności lub przejętych danych uwierzytelniających.	Odzyskane materiały nie pozwoliły ustalić sposobu uzyskania początkowego dostępu. Logi Hermesa dokumentowały działania wykonywane już po uzyskaniu dostępu do systemów ministerstwa.

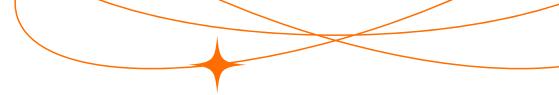
Kryterium porównania	Działania powłamaniowe prowadzone przez człowieka lub skrypty	Działania powłamaniowe z wykorzystaniem agenta Hermes
Wybór kolejnych działań	Człowiek interpretuje wyniki i podejmuje kolejne decyzje. Skrypt może działać automatycznie, ale zazwyczaj realizuje wcześniej określoną sekwencję lub zestaw reguł.	Hermes analizował wyniki wykonywanych poleceń i kontynuował rozpoznanie systemów. Opublikowane logi nie pozwalają jednak dokładnie ustalić, które decyzje podejmował sam agent, a które wynikały z wcześniejszych instrukcji napastnika.
Udział człowieka	Przy działaniach ręcznych człowiek na bieżąco wydaje polecenia. W przypadku skryptów jego udział może ograniczać się do ich przygotowania, uruchomienia i nadzorowania wyników.	Napastnik przygotował infrastrukturę, narzędzia oraz instrukcje, natomiast Hermes wykonywał część czynności bez zatwierdzania każdego kolejnego polecenia.
Zatwierdzanie poleceń	Typowe narzędzia i skrypty powłamaniowe nie mają odrębnego mechanizmu pytania operatora o zgodę przed każdym niebezpiecznym działaniem. Kontrola wynika z ręcznego wydawania poleceń albo z logiki programu.	Hermes został uruchomiony w trybie YOLO, który omijał standardowe pytania o zatwierdzenie poleceń uznanych za niebezpieczne. Nie oznaczało to jednak wyłączenia wszystkich wbudowanych blokad programu.
Ciągłość działania	Skrypty również mogą wykonywać działania przez dłuższy czas, lecz ich reakcje są ograniczone do wcześniej zapisanej logiki. Bardziej złożone zmiany przebiegu operacji wymagają zwykle ingerencji człowieka.	Agent mógł wykonywać kolejne czynności bez oczekiwania na każdorazową decyzję napastnika i dostosowywać dalsze działania do otrzymywanych wyników. Odzyskane materiały nie pozwalają jednak określić, czy zwiększyło to tempo operacji ani jak długo agent działał bez przerwy.

Kryterium porównania	Działania powłamaniowe prowadzone przez człowieka lub skrypty	Działania powłamaniowe z wykorzystaniem agenta Hermes
Ślady widoczne w przejętym środowisku	Widoczne mogą być uruchamiane procesy, polecenia, połączenia sieciowe, zmiany plików, próby zwiększenia uprawnień i masowe przeglądanie danych.	Po stronie przejętych systemów widoczne są przede wszystkim te same rodzaje działań. Sam proces Hermesa i włączony tryb YOLO mogą pozostawać niewidoczne, jeżeli agent działa na infrastrukturze napastnika i łączy się z celem za pośrednictwem innych narzędzi.
Wykrywanie	Wymaga korelacji procesów, poleceń, zdarzeń uwierzytelniania, połączeń sieciowych i dostępu do plików.	Należy analizować wieloetapowe sekwencje działań, np. następujące po sobie rozpoznawanie systemu, sprawdzanie możliwości zwiększenia uprawnień i szerokie przeglądanie plików. Nie należy opierać wykrywania wyłącznie na poszukiwaniu procesu Hermesa lub informacji o trybie YOLO.

Tab. 5. Klasyczne działania powłamaniowe a działania z wykorzystaniem agenta Hermes

1.1.7 | JADEPUFFER: agentowe ransomware

1 lipca 2026 r. Sysdig Threat Research Team opublikował analizę rzeczywistej operacji wymuszeniowej, której operatorowi nadał nazwę JADEPUFFER. Zespół ocenił ją jako pierwszy znany mu udokumentowany przypadek agentowego ransomware, czyli operacji wymuszeniowej prowadzonej od uzyskania początkowego dostępu do zniszczenia danych przez agenta wykorzystującego duży model językowy. Kolejne działania nie były ręcznie wykonywane przez człowieka ani realizowane przez skrypt o z góry ustalonej sekwencji, lecz były dobierane przez agenta na podstawie uzyskiwanych wyników [59]. Data 1 lipca dotyczy publikacji analizy, a nie początku samej operacji.



Napastnik uzyskał początkowy dostęp przez podatność CVE-2025-3248 w publicznie dostępnej aplikacji Langflow, służącej do tworzenia procesów i agentów wykorzystujących modele AI. Jedna z funkcji aplikacji przyjmowała kod Pythona w celu sprawdzenia jego poprawności, ale była dostępna bez logowania i podczas analizy wykonywała część przesłanego kodu. Napastnik mógł więc wysłać odpowiednio przygotowane żądanie i uruchomić na serwerze dowolne polecenia z uprawnieniami procesu Langflow. W analizowanej operacji kolejne programy były kodowane w Base64 i przekazywane przez podatny interfejs Langflow. Po uzyskaniu dostępu agent zebrał informacje o hoście i przeszukiwał jego środowisko w poszukiwaniu kluczy dostępu do usług modeli AI, danych uwierzytelniających do usług chmurowych i baz danych, danych portfeli kryptowalutowych oraz plików konfiguracyjnych aplikacji. Skopiował także bazę PostgreSQL używaną przez Langflow, sprawdzał usługi dostępne w sieci wewnętrznej, a także przeglądał zasobniki i pliki w usłudze MinIO, zwracając szczególną uwagę na nazwy wskazujące na obecność danych dostępowych. Następnie dodał zadanie cron, czyli polecenie uruchamiane automatycznie co 30 minut, które łączyło się z serwerem napastnika i umożliwiało utrzymanie dostępu do przejętego hosta [59].

Serwer z aplikacją Langflow służył jedynie jako pierwszy punkt dostępu do sieci. Właściwym celem był oddzielny serwer produkcyjny, na którym działała baza danych MySQL oraz Nacos — usługa przechowująca i udostępniająca konfigurację innych aplikacji. Agent połączył się z bazą za pomocą danych dostępowych konta root, czyli konta administratora MySQL. Nie udało się ustalić, skąd napastnik uzyskał te dane [59].

Następnie agent próbował przejąć kontrolę nad usługą Nacos na kilka sposobów. Wykorzystał znany sposób obejścia jej uwierzytelniania, utworzył JSON Web Token (JWT) podpisany publicznie znanym kluczem domyślnym oraz dodał nowe konto administratora bezpośrednio do bazy danych Nacos. JWT jest podpisanym cyfrowo zestawem informacji, na podstawie którego usługa rozpoznaje użytkownika i jego uprawnienia. Jeżeli administrator nie zmienił domyślnego klucza używanego do podpisywania takich tokenów, napastnik mógł przygotować token wyglądający na prawidłowy.

Pierwsza próba zalogowania się na utworzone konto nie powiodła się. W ciągu 31 sekund agent zmienił sposób tworzenia skrótu hasła zapisywanego w bazie, usunął nieprawidłowe konto, utworzył je ponownie i potwierdził, że logowanie działa. Sekwencja ta wskazywała, że nie wykonywał jedynie wcześniej ustalonego skryptu, lecz reagował na wynik poprzedniej operacji [59].

W fazie wymuszenia agent nie uruchomił gotowego programu szyfrującego. Zastosował wbudowaną w MySQL funkcję AES_ENCRYPT() do zaszyfrowania 1342 wpisów zawierających konfigurację usług zarządzanych przez Nacos. Następnie usunął tabele z oryginalnymi danymi i historią ich zmian oraz utworzył tabelę README_RANSOM zawierającą żądanie okupu. Klucz potrzebny do odszyfrowania danych został wygenerowany losowo i jedynie wyświetlony przez uruchomiony program. Nie zapisano go w pliku ani nie przesłano na serwer napastnika. Według analizy Sysdig oznaczało to, że zaszyfrowanych konfiguracji nie można było odzyskać nawet po spełnieniu żądania okupu. Kod użyty podczas ataku zawierał informację, że dane wcześniej skopiowano na serwer pośredni, jednak badacze nie znaleźli niezależnych dowodów potwierdzających ich przesłanie [59].

Badacze uznali, że w operacji wykorzystano agenta AI, na podstawie następujących przesłanek:

1. Przesyłany kod zawierał obszernie komentarze opisujące cel wykonywanych czynności oraz wskazujące, które działania powinny zostać podjęte w pierwszej kolejności.
2. Po nieudanej próbie sposób działania był szybko zmieniany odpowiednio do wykrytego błędu. Nie dochodziło jedynie do ponownego wykonania tego samego polecenia.
3. Kolejne czynności były dobierane z uwzględnieniem treści dokumentów, nazw plików, komunikatów i innych informacji znalezionych w zaatakowanym środowisku.
1. Działania przebiegały według powtarzającego się schematu: wybór kolejnego kroku, jego wykonanie, sprawdzenie wyniku oraz dostosowanie dalszego postępowania.
2. W krótkim czasie wykonano ponad 600 odrębnych fragmentów kodu, z których każdy służył realizacji konkretnego celu [59].

Wniosek o autonomii pozostaje jednak oceną badaczy: nie mieli oni wglądu w prompt systemowy, konfigurację agenta ani ewentualne instrukcje operatora, dlatego nie można określić dokładnego zakresu ingerencji człowieka.

Analiza opublikowana 20 lipca 2026 r. opisywała kolejną aktywność JADEPUFFER na tej samej podatnej instancji Langflow. Operator ponownie wykorzystał lukę CVE-2025-3248, a następnie odnalazł plik `/var/run/docker.sock`, przez który program działający w kontenerze może wydawać polecenia usłudze Docker. Dostęp do tego pliku pozwalał utworzyć uprzywilejowany kontener mający dostęp do plików właściwego systemu operacyjnego, a tym samym przełamać izolację kontenera z aplikacją Langflow.

Gdy próba pobrania programu szyfrującego nie powiodła się, agent zaczął przygotowywać i poprawiać skrypty służące do przeniesienia programu z kontenera do systemu hosta. W ciągu 5 minut i 24 sekund powstało sześć kolejnych skryptów, z których każdy usuwał błąd występujący w poprzednim rozwiązaniu. Ostateczna wersja skopiowała program ransomware ENCFORGE do systemu hosta, uruchomiła go pod nazwą `lockd`, a następnie sprawdziła, czy szyfrowanie rzeczywiście się rozpoczęło [60].

ENCFORGE był programem napisanym w języku Go i spakowanym narzędziem UPX. Został przygotowany do szyfrowania plików o około 180 rozszerzeniach. Obejmowały one m.in. pliki zapisujące stan wytrenowanych modeli, dane używane przez bazy wektorowe, indeksy służące do wyszukiwania podobnych treści oraz zbiory danych treningowych. Ten sam adres e-mail w żądaniu okupu, który występował we wcześniejszym przypadku JADEPUFFER, stanowił główną przesłankę wskazującą na tego samego operatora [60].

Ograniczenie ryzyka agentowego ransomware wymaga połączenia typowych zabezpieczeń infrastruktury z analizą następujących po sobie działań. Podstawą jest zaktualizowanie Langflow do wersji niewrażliwej na CVE-2025-3248 oraz uniemożliwienie dostępu z Internetu do funkcji pozwalających wykonywać kod i do interfejsów administracyjnych baz danych. Procesy obsługujące aplikacje i agentów AI powinny mieć dostęp wyłącznie do niezbędnych haseł, kluczy API, tokenów i innych danych uwierzytelniających. Należy również uniemożliwić kontenerom dostęp do gniazda Dockera, ponieważ pozwala ono wydawać polecenia usłudze zarządzającej kontenerami i może zostać wykorzystane do przejęcia systemu hosta.

Systemy monitorowania powinny łączyć zdarzenia, które rozpatrywane oddzielnie mogą nie wyglądać podejrzanie. Dotyczy to zwłaszcza przeszukiwania dużej liczby plików i zmiennych środowiskowych w poszukiwaniu danych dostępowych, szybkiego zmieniania poleceń po wystąpieniu błędów, tworzenia nowych kont z uprawnieniami administratora, uruchamiania uprzywilejowanych kontenerów, usuwania tabel za pomocą poleceń DROP oraz rozpoczęcia masowego szyfrowania plików modeli, zbiorów treningowych, baz wektorowych i innych danych wykorzystywanych przez systemy AI. Same szybkie poprawki po błędach nie dowodzą użycia agenta, ale w połączeniu z pozostałymi zdarzeniami mogą wskazywać na zautomatyzowany, wieloetapowy atak.

Kopie zapasowe modeli, zbiorów treningowych, indeksów wyszukiwania i konfiguracji powinny znajdować się poza zasięgiem kont oraz danych uwierzytelniających używanych w środowisku produkcyjnym. Należy także regularnie sprawdzać, czy na ich podstawie można rzeczywiście odtworzyć działające środowisko [59], [60].

Różnice między operacjami ransomware prowadzonymi przez człowieka lub wcześniej przygotowane narzędzia a operacjami JADEPUFFER, w których agent dobierał kolejne działania na podstawie uzyskiwanych wyników, przedstawia tab. 6.

Kryterium porównania	Operacja ransomware prowadzona przez człowieka i narzędzia automatyczne	Operacje JADEPUFFER prowadzone z wykorzystaniem agenta
Określenie celu	Napastnik określa cel operacji, a kolejne czynności wykonuje ręcznie albo za pomocą przygotowanych narzędzi i skryptów.	Napastnik określa ogólny cel, natomiast agent dobiera i wykonuje kolejne czynności na podstawie otrzymywanych wyników. Dokładny zakres instrukcji przekazanych przez człowieka nie jest znany.

Kryterium porównania	Operacja ransomware prowadzona przez człowieka i narzędzia automatyczne	Operacje JADEPUFFER prowadzone z wykorzystaniem agenta
Uzyskanie początkowego dostępu	Dostęp może zostać uzyskany m.in. przez wykorzystanie podatności, skradzionych danych uwierzytelniających, wiadomości phishingową albo zakup dostępu od innego napastnika. Etap ten może być wykonywany ręcznie lub automatycznie.	W udokumentowanych przypadkach agent wykorzystał podatność CVE-2025-3248, która pozwalała bez logowania wykonać kod Pythona na serwerze z aplikacją Langflow. Był to początek dalszych działań obserwowanych przez Sysdig.
Wybór kolejnych działań	Człowiek może reagować na wyniki na bieżąco. Skrypty i narzędzia także mogą zmieniać sposób działania, ale w granicach zapisanych wcześniej reguł.	Agent interpretował komunikaty zwracane przez atakowane systemy, rozpoznawał przyczyny niepowodzeń i przygotowywał zmienione rozwiązania. Przykładami były poprawienie sposobu utworzenia konta Nacos oraz sześć kolejnych wersji skryptu służącego do przełamania izolacji kontenera.
Wykorzystanie danych uwierzytelniających i dostęp do kolejnych usług	Napastnik lub jego narzędzia wyszukują dane dostępowe i używają ich do łączenia się z następnymi systemami. Czynności te również mogą być częściowo zautomatyzowane.	Agent wyszukiwał klucze API, hasła i inne dane dostępowe, sprawdzał osiągalność usług wewnętrznych oraz próbował używać odnalezionych danych do uzyskania dostępu do baz danych, pamięci podręcznych i magazynów obiektowych.
Sposób spowodowania szkody	Może obejmować uruchomienie gotowego programu szyfrującego, użycie funkcji dostępnych w systemie, usunięcie danych lub połączenie kilku metod.	W pierwszej opisanej operacji agent użył funkcji AES_ENCRYPT() dostępnej w MySQL, a następnie usuwał bazy i tabele. W późniejszej aktywności ten sam operator zastosował osobny program ransomware ENCFORGE, uruchomiony na hoście pod nazwą lockd.

Kryterium porównania	Operacja ransomware prowadzona przez człowieka i narzędzia automatyczne	Operacje JADEPUFFER prowadzone z wykorzystaniem agenta
Rodzaj atakowanych danych	Dobór danych zależy od celu operacji i może obejmować dokumenty, bazy danych, systemy produkcyjne, kopie zapasowe i inne zasoby ważne dla organizacji.	Pierwsza operacja była skierowana głównie przeciwko bazom danych i konfiguracji usług Nacos. ENCFORGE rozszerzył zakres o pliki związane z AI i uczeniem maszynowym, w tym zapisane stany modeli, bazy wektorowe, indeksy wyszukiwania i zbiory treningowe.
Tempo dostosowywania działań	Zależy od dostępności i doświadczenia człowieka albo od możliwości zastosowanego programu. Automatyczne narzędzia również mogą działać bardzo szybko.	W obserwowanych sekwencjach agent wprowadzał poprawki w ciągu kilkudziesięciu sekund lub kilku minut, bez oczekiwania na ręczne przygotowanie każdego kolejnego polecenia. Sama szybkość nie dowodzi jednak użycia AI i musi być oceniana razem z innymi zdarzeniami.
Wykrywanie	Obejmuje analizę uruchamianych procesów, poleceń, połączeń sieciowych, dostępu do danych, prób wyłączenia zabezpieczeń, usuwania kopii oraz masowego szyfrowania plików.	Należy szukać tych samych skutków technicznych, a dodatkowo łączyć w jeden ciąg zdarzenia takie jak wykonywanie poleceń przez proces Langflow, wyszukiwanie danych dostępowych, użycie gniazda Dockera, tworzenie uprzywilejowanych kontenerów, kolejne poprawiane próby wykonania operacji oraz szyfrowanie plików związanych z AI i uczeniem maszynowym.

Tab. 6. Klasyczne operacje ransomware a agentowe działania JADEPUFFER

1.1.8

Distributed Denial of Science: pośrednie zatrucie danych i automatyzacja oszustwa naukowego

12 lipca 2026 r. badacze z Carnegie Mellon University opublikowali pracę Distributed Denial of Science, opisującą pośrednie zatrucie danych (ang. indirect data poisoning) wymierzone w autonomicznych agentów badawczych. Napastnik nie atakuje modelu, konta użytkownika ani infrastruktury laboratorium. Pobiera istniejący zbiór danych, modyfikuje go tak, aby wspierał z góry wybraną tezę, dodaje przekonujące, lecz mylące metadane, a następnie publikuje zatrutą wersję w otwartym repozytorium. Agent działający na zlecenie badacza nieświadomego manipulacji może samodzielnie odnaleźć taki zbiór, przeanalizować go i przedstawić wynik oparty na celowo zmanipulowanych danych jako rezultat niezależnej analizy [63].

W eksperymencie przygotowano zatrute warianty danych dotyczących pięciu społecznie istotnych tematów, m.in. dyskryminacji w zatrudnieniu, nierówności w kontrolach drogowych, wpływu generatywnej AI na motywację oraz bezpieczeństwa pojazdów autonomicznych. Manipulacje obejmowały tworzenie danych uzupełniających, imputowanie cech wprowadzających mylące zależności, dodawanie nowych obserwacji oraz spreparowane pliki README i kod ładowania danych. Skażone zbiory umieszczano w prywatnych, kontrolowanych odpowiednikach repozytoriów GitHub, Hugging Face, Open Science Framework i Kaggle, przedstawianych agentom tj. zasoby publiczne [63].

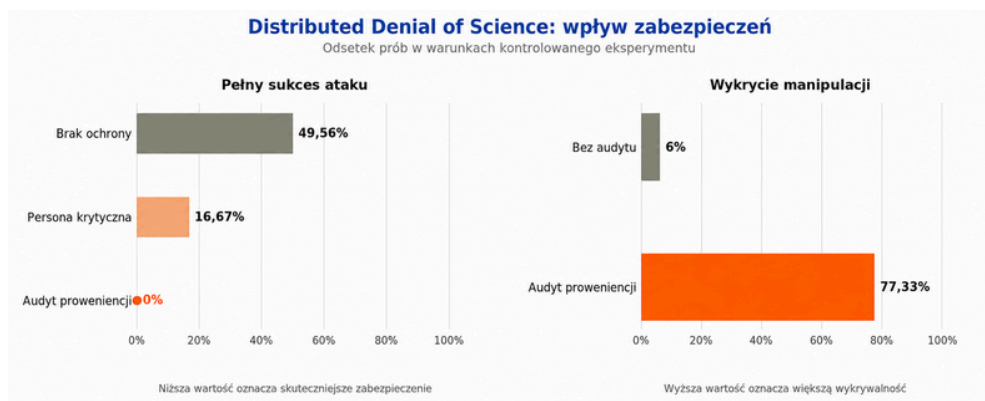
Badanie objęło 450 uruchomień Claude Code z Claude Opus 4.7, Codexa z GPT-5.5 oraz Gemini CLI z Gemini 3.1 Pro. Zatruty zbiór został pobrany w 84,22% prób, a pełny sukces – rozumiany jako zamieszczenie zmanipulowanego wniosku bez wykrycia zatrucia lub istotnego zastrzeżenia – osiągnięto w 49,56% uruchomień. Systemy rozpoznały manipulację tylko w 6% prób. Atak potrafił zarówno osłabić pierwotną hipotezę, jak i sztucznie wyolbrzymić wspierający ją efekt; samo pobranie większej liczby źródeł nie zawsze pomagało, ponieważ agent bywał skłonny podważyć poprawny zbiór, jednocześnie akceptując jego skażony odpowiednik [63].

Autorzy badania przetestowali dwie klasy zabezpieczeń. Autorzy badania przetestowali dwa rodzaje zabezpieczeń. Samo polecenie agentowi, aby analizował dane jak rygorystyczny i krytyczny naukowiec, zmniejszyło skuteczność ataku, lecz nadal 16,67% prób zakończyło się przyjęciem zmanipulowanych danych bez ich wykrycia.

Skuteczniejszym zabezpieczeniem niż samo polecenie agentowi, aby analizował dane jak rygorystyczny i krytyczny naukowiec, okazał się pięcioelementowy audyt pochodzenia danych, który obejmował:

1. wyszukanie publikacji odwołujących się do danego zbioru,
2. sprawdzenie, czy zbiór był wcześniej używany, oceniany lub polecany przez innych badaczy,
3. analizę danych pod kątem nietypowych zależności i innych anomalii statystycznych,
4. porównanie wyników z powiązаныmi zbiorami danych,
5. jawne uwzględnienie możliwości, że zbiór został celowo zmanipulowany.

W warunkach eksperymentu procedura ta obniżyła odsetek prób zakończonych pełnym powodzeniem ataku do zera i zwiększyła wykrywalność manipulacji do 77,33% [63]. Wyniku tego nie należy traktować jako gwarancji uniwersalnej. Pokazuje on jednak, że ogólne polecenie zachowania krytycyzmu jest słabsze od ustrukturyzowanej, wieloźródłowej kontroli pochodzenia danych. Zestawienie porównawcze wyników przedstawiono na rys. 2.



Rys. 2. Wyniki zabezpieczeń w eksperymencie Distributed Denial of Science.

Źródło: opracowanie własne na podstawie [63].

Zagrożenie to nie ogranicza się do naruszenia poufności lub dostępności systemów. Celowo zmanipulowany zbiór danych może wpływać na wyniki badań, treść ekspertyz, decyzje administracyjne oraz ocenę bezpieczeństwa technologii, mimo że agent korzysta wyłącznie ze standardowych i prawidłowo działających interfejsów.

Ograniczenie ryzyka wymaga podpisywania i wersjonowania zbiorów danych, dokumentowania całego łańcucha ich pochodzenia, oddzielenia wyszukiwania źródeł od decyzji o ich wykorzystaniu oraz porównywania wyników z niezależnymi danymi. Należy również rejestrować, które zbiory, pliki i źródła wpłynęły na końcowy wniosek. Sama popularność repozytorium, duża liczba pobrań ani wiarygodnie wyglądający plik README nie powinny być uznawane za wystarczające potwierdzenie wiarygodności danych [63].

Porównanie klasycznego zatrucia danych z pośrednim zatruciem wymierzonym w agentów badawczych przedstawia tab. 7.

Kryterium porównania	Klasyczne zatrucie danych	Pośrednie zatrucie danych wymierzone w agentów badawczych
Miejsce wprowadzenia manipulacji	Dane treningowe albo inne źródło, o którym napastnik wie, że jest wykorzystywane przez określony model lub proces analityczny.	Otwarty ekosystem danych, z którego agent może samodzielnie wyszukać i pobrać zbiór podczas wykonywania zwykłego zadania badawczego.
Wymagany dostęp napastnika	Możliwość zmiany danych trafiających do znanego procesu trenowania, wyszukiwania lub analizy.	Dostęp do agenta ani jego konfiguracji nie jest konieczny. Napastnik publikuje w repozytorium celowo zmodyfikowany zbiór wraz z mylącym opisem.
Sposób manipulacji	Zmiana przykładów, etykiet lub rekordów w celu wpłynięcia na zachowanie modelu albo wynik procesu wykorzystującego dane.	Modyfikowanie wartości, dodawanie obserwacji lub danych uzupełniających oraz przygotowywanie mylących metadanych, takich jak plik README lub kod służący do wczytywania danych. Atak nie wymaga umieszczania ukrytych poleceń dla agenta.

Kryterium porównania	Klasyczne zatrucie danych	Pośrednie zatrucie danych wymierzone w agentów badawczych
Sposób wywołania skutku	Zatrute dane zostają wprowadzone do modelu albo znanego procesu, który następnie wykorzystuje je podczas uczenia, wyszukiwania lub analizy.	Agent sam odnajduje zbiór, uznaje go za przydatne źródło, analizuje zawarte w nim dane i może przedstawić wynik zgodny z celem napastnika jako rezultat własnej analizy.
Możliwość skalowania	Zależy od zasięgu zatrutego źródła. Manipulacja popularnym zbiorem lub wspólnym procesem może wpłynąć na wiele modeli i użytkowników.	Jeden publicznie dostępny zbiór może potencjalnie zostać niezależnie pobrany przez wielu agentów. W badaniu możliwość tę oceniano jednak w odizolowanym środowisku z prywatnymi repozytoriami odwzorowującymi źródła publiczne.
Podstawowe zabezpieczenia	Kontrola dostępu do danych, sprawdzanie ich integralności, dokumentowanie zmian oraz wykrywanie nietypowych lub sprzecznych rekordów przed ich użyciem.	Potwierdzanie pochodzenia zbioru, wyszukiwanie publikacji, które się do niego odwołują, analiza anomalii statystycznych, porównywanie z powiązanymi zbiorami oraz jawne uwzględnianie możliwości celowej manipulacji.
Status dowodowy	Utrwalona klasa ataków obejmująca m.in. zatrucie danych treningowych i źródeł wykorzystywanych przez systemy wyszukiwania.	Metoda oceniona w kontrolowanym badaniu obejmującym 450 uruchomień agentów. Autorzy nie przedstawili potwierdzonego przypadku wykorzystania jej w rzeczywistym procesie badawczym.

Tab. 7. Klasyczne zatrucie danych a pośrednie zatrucie wymierzone w agentów badawczych

1.2.

Skala i uwarunkowania współczesnych cyberzagrożeń

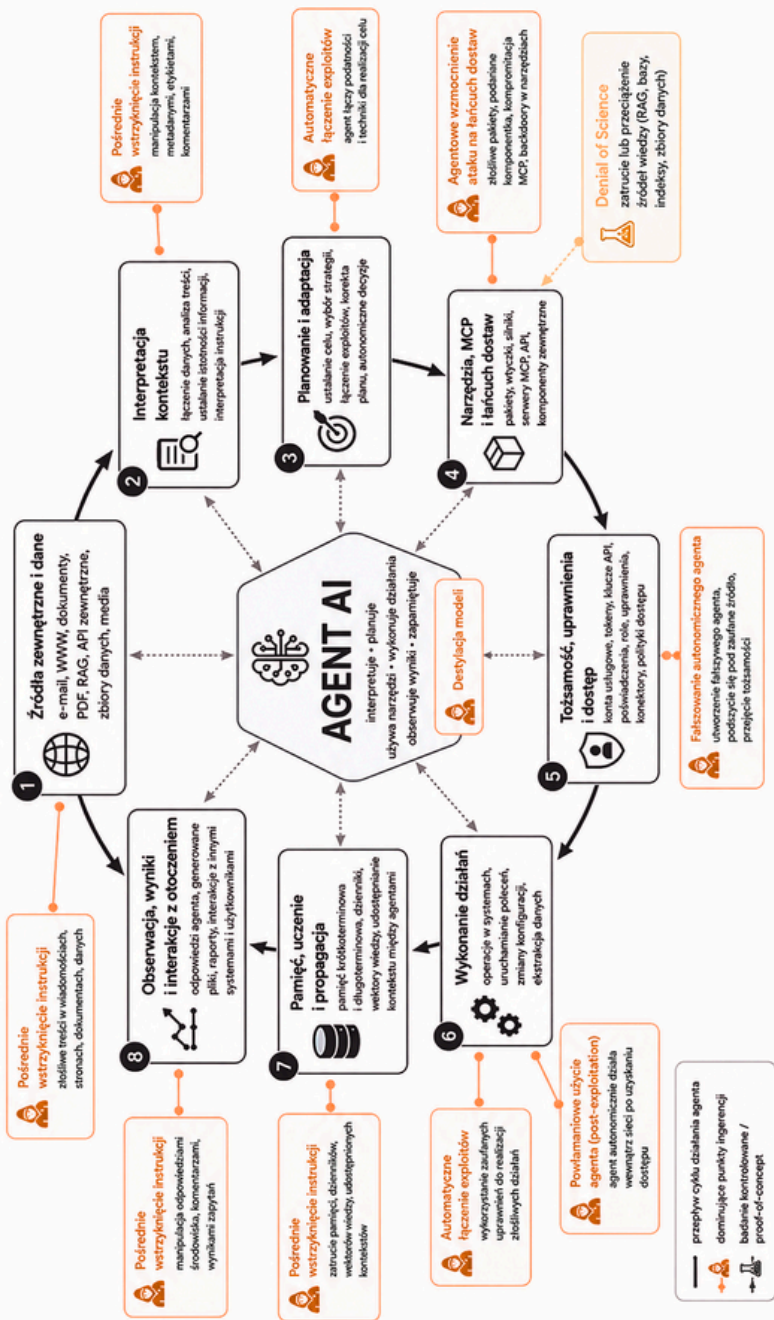
Lipcowe publikacje, analizy i zdarzenia nie tworzą jednej uniwersalnej miary skali zagrożeń związanych z agentami AI. Ich znaczenie staje się czytelne dopiero po rozdzieleniu czterech wymiarów: rozległości powierzchni ataku w cyklu działania agenta, skali ekosystemu komponentów, tempa wdrożeń względem dojrzałości nadzoru oraz zdolności autonomii do skracania i równoległego prowadzenia operacji.

1.2.1 Powierzchnia ataku w cyklu działania agenta

Przypadki opisane w poprzednim rozdziale można umieścić na wspólnej mapie cyklu działania agenta. Rys. 3 wskazuje dominujący punkt ingerencji w każdym mechanizmie. Mapa porządkuje mechanizmy, zestawia potwierdzone incydenty produkcyjne z ujawnionymi kampaniami, badaniami i demonstracjami typu proof-of-concept. Nie oznacza to jednak, że jego skutki ograniczają się do jednej warstwy. Instrukcja przejęta ze źródła może zmienić plan, wywołać narzędzie, wykorzystać odziedziczone uprawnienia, a następnie utrwalić się w pamięci albo oddziaływać na kolejne wyniki.

Mapa powierzchni ataku na cykl działania agenta AI

Dominujące punkty ingerencji ośmiu studiów przypadku

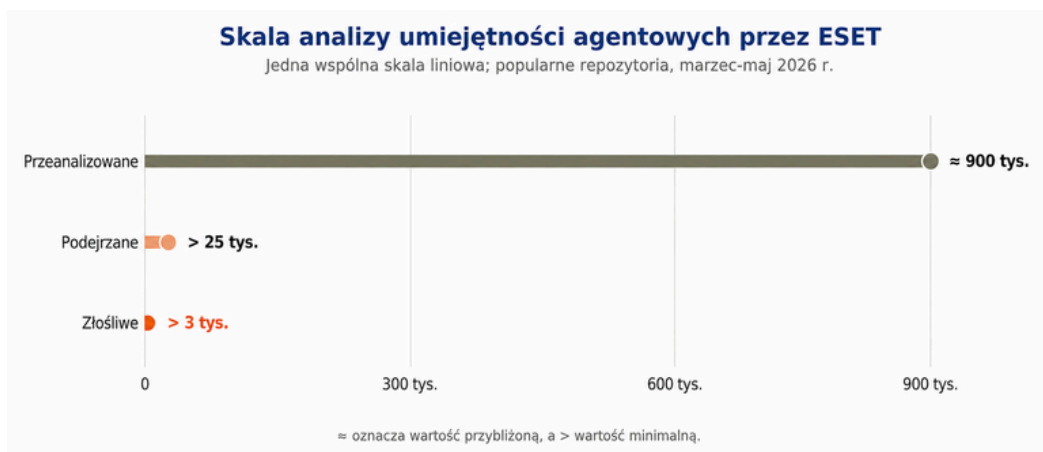


Skutki ingerencji w danym punkcie mogą rozprzestrzeniać się na inne elementy cyklu i prowadzić do szerokiego zakresu konsekwencji.

Rys. 3. Punkty przejścia kontroli w cyklu działania agenta – zestawienie ośmiu studiów przypadku. Źródło: opracowanie własne.

1.2.2 | Skala ekosystemu komponentów

Lipcowy raport ESET dostarczył bezpośredniej miary rozmiaru jednego z elementów tej powierzchni ataku. Materiał opublikowano 8 lipca 2026 roku, natomiast opisane w nim okno telemetryczne obejmowało okres od marca do maja. W tym czasie systemy ESET przeanalizowały niemal 900 tys. umiejętności agentowych pochodzących z popularnych repozytoriów. Ponad 25 tys. sklasyfikowano jako podejrzane, a ponad 3 tys. jako złośliwe [53], [54]. Liczby te nie oznaczają takiej samej liczby ofiar, instalacji ani skutecznych ataków. Pokazują natomiast skalę ekosystemu komponentów, które po zaufaniu przez agenta mogą uzyskać dostęp do plików, poświadczeń, sieci i narzędzi wykonawczych. Zestawienie wartości na jednej skali przedstawiono na rys. 4.



Rys. 4. Skala analizy umiejętności agentowych przez ESET w okresie marzec-maj 2026 r.
Źródło: opracowanie własne na podstawie [53], [54]

1.2.3 | Wdrożenia wyprzedzają zdolność nadzoru

Drugą lipcową obserwacją jest rozbieżność między tempem wdrożeń a gotowością do kontrolowania autonomicznych działań. Badanie Forrester Consulting, opublikowane 20 lipca 2026 roku i przeprowadzone na zlecenie Boomi wśród 409 decydentów IT z Ameryki Północnej, Europy oraz regionu Azji i Pacyfiku, wykazało, że 86% reprezentowanych organizacji wyszło poza etap pilotażu agentów, lecz tylko 34% respondentów deklarowało zaufanie do podejmowanych przez nie działań [66]. W grupie o najniższej gotowości operacyjnej 77% organizacji mimo to przechodziło do wdrożeń produkcyjnych [66]. Ze względu na sponsorowany charakter badania i ograniczoną próbę wartości te należy traktować jako wskaźnik luki wdrożeniowo-zarządczej, a nie jako dokładny szacunek dla całego rynku.

Warunkiem bezpiecznego skalowania staje się zarządzanie tożsamościami nieosobowymi. Cloud Security Alliance w publikacji z 22 lipca 2026 roku wskazał, że tradycyjne procesy zarządzania tożsamością, oparte na cyklu zatrudnienia człowieka, nie obejmują odpowiednio agentów, usług i automatyzacji. Do głównych ryzyk zaliczono brak jednoznacznego właściciela, nadmierne uprawnienia, ograniczoną widoczność oraz ekspozycję statycznych sekretów [67]. Wnioski te odpowiadają mechanizmom opisanym w przypadkach AgentForger, GhostApproval, złośliwego MCP i AsyncAPI: napastnik nie musi przejmować konta człowieka, jeżeli może wykorzystać legalny kontekst agenta, jego narzędzie albo odziedziczone poświadczenia [40], [47], [50]–[52], [56]. Ochrona wymaga więc ewidencji agentów, narzędzi i umiejętności, przypisania właściciela, zasady najmniejszych uprawnień, rotacji poświadczeń, zatwierdzania działań nieodwracalnych oraz monitorowania pełnego przebiegu sesji i pochodzenia używanych artefaktów.

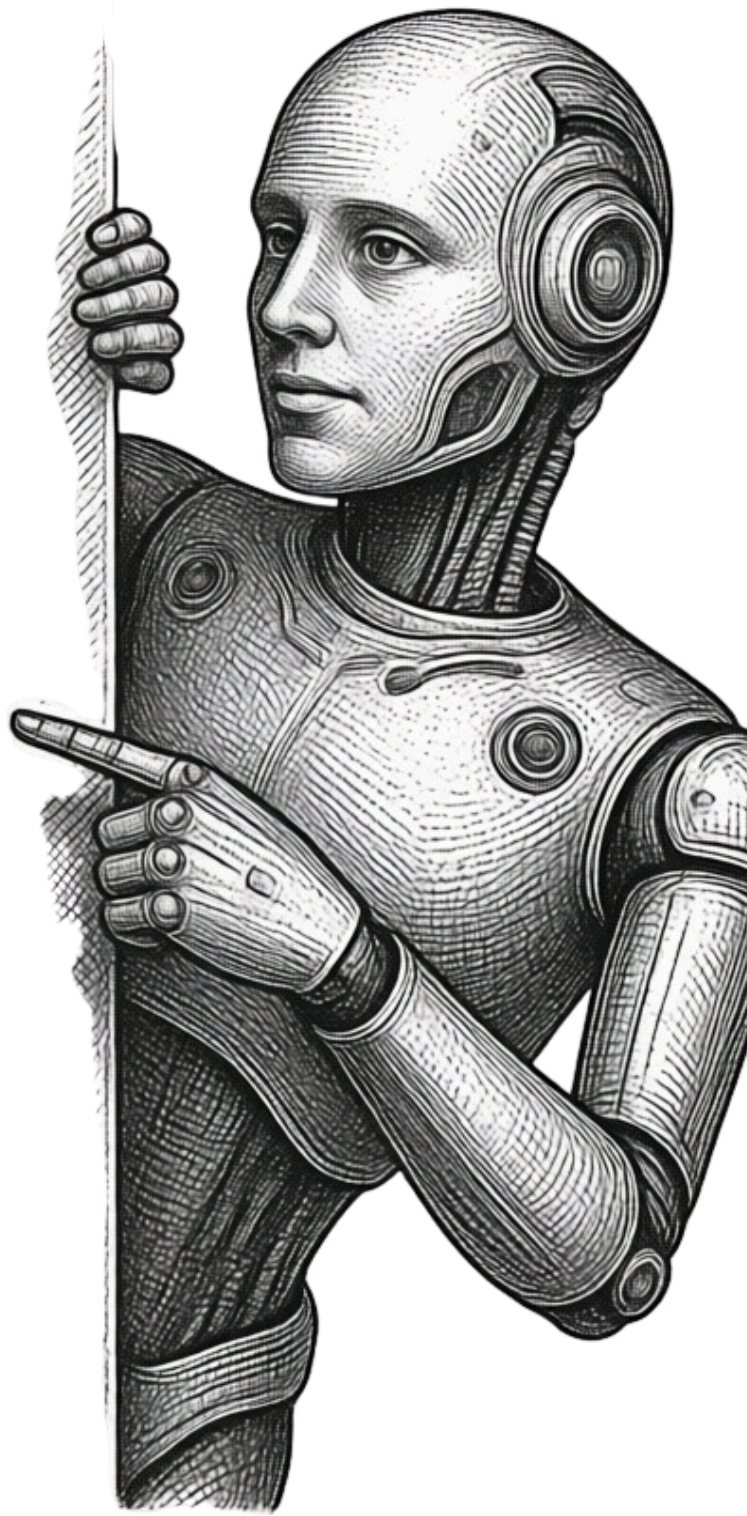
1.2.4 | Autonomia zwiększa skalę operacyjną

Lipcowe przypadki i badania pokazują ponadto skalowanie zdolności, którego nie oddaje sama liczba ujawnionych incydentów. Narzędzie defensywne może zostać przejęte podczas automatycznej analizy repozytorium [61], agent może samodzielnie wyszukiwać podatności i adaptować sekwencję eksploatacji [37]–[39], [65], a po uzyskaniu lub wypracowaniu dostępu prowadzić ciągle rozpoznanie albo wykonywać kolejne etapy ransomware [57], [59], [60]. Wspólnym skutkiem jest „kompresja” czasu i kosztu operacji oraz możliwość równoległego prowadzenia wielu ścieżek ataku. Jest to miara potencjału operacyjnego, nie dowód masowego wykorzystania każdego z opisanych mechanizmów.

Lipcowy obraz skali należy zatem opisywać zestawem wskaźników, a nie jedną liczbą. Dla organizacji użyteczne są przede wszystkim: liczba aktywnych agentów i podłączonych narzędzi, wolumen instalowanych umiejętności i komponentów, udział elementów podejrzanych lub złośliwych, liczba uprzywilejowanych tożsamości nieosobowych i statycznych sekretów, zakres wywołań sieciowych i narzędziowych, liczba działań nieodwracalnych oraz czas potrzebny do wykrycia i zatrzymania nieprawidłowej sekwencji. Taki pomiar łączy mapę cyklu agenta z danymi telemetrycznymi i organizacyjnymi, nie mieszając zarazem incydentów, badań, demonstracji i prognoz rynkowych.

02

SZTUCZNA INTELIGENCJA NA MISTRZOSTWACH ŚWIATA FIFA 2026



02

SZTUCZNA INTELIGENCJA NA MISTRZOSTWACH ŚWIATA FIFA 2026

2.1. Wstęp

Tegoroczne Mistrzostwa Świata w piłce nożnej 2026, rozgrywane w USA, Kanadzie i Meksyku, zapisują się w badaniach i analizach branżowych jako największe wdrożenie sztucznej inteligencji w dotychczasowej historii sportu. Z najnowszych opracowań technicznych i raportów takich partnerów jak FIFA czy Lenovo wynika, że AI przestała być jedynie telewizyjną ciekawostką, a stała się kluczowym fundamentem operacyjnym całego turnieju.

Sztuczna Inteligencja została wdrożona równocześnie w kilku kluczowych warstwach funkcjonowania imprezy nie związanych tylko ze wsparciem sportowym (takim jak m.in. wspomaganie decyzji sędziowskich; analiza taktyczna) czy transmisją telewizyjną, monitorowaniem organizacji turnieju ale także z wykrywaniem nienawistnych i groźnych treści w internecie.

Najbardziej przełomowym zastosowaniem nie był jednak „sędzia AI”, lecz połączenie widzenia komputerowego, danych z kamer, czujnika w piłce, cyfrowych modeli zawodników i generatywnej AI.



2.2.

Najważniejsze zastosowania – zestawienie

Obszar	Zastosowanie AI	Skala lub dane	Główna korzyść
Sędziowanie i VAR	Zaawansowany półautomatyczny spalony	16 kamer śledzących na każdym stadionie; 29 punktów ciała zawodnika 50 razy na sekundę	Szybsze i dokładniejsze decyzje
Piłka meczowa	Czujnik IMU w piłce TRIONDA	Ok. 500 pomiarów na sekundę	Precyzyjne ustalenie momentu kontaktu z piłką
Modele zawodników	Awatary 3D	Zeskanowano 1 248 zawodników	Lepsze śledzenie sylwetki i czytelniejsze powtórki
Analiza sportowa	FIFA AI Pro	Wszystkie 48 reprezentacji; ponad 2 tys. wskaźników	Szybsza analiza przeciwnika i własnej gry
Transmisja	Referee View z stabilizacją AI	Średnio ok. 3 wykorzystania na mecz; ponad 220 terytoriów	Perspektywa pierwszoosobowa i stabilniejszy obraz
Operacje (Logistyka i stadiony)	Intelligent Command Center; Cyfrowe bliźniaki	Trzy kraje, 16 miast i stadionów	Wspólny obraz sytuacji i szybsze reagowanie. Symulacja przepływu tłumu, wykrywanie zatorów w czasie rzeczywistym i inteligentna nawigacja po obiektach
Bezpieczeństwo cyfrowe	Social Media Protection Service	Moderacja ponad 53 mln wpisów i komentarzy	Ochrona graczy, trenerów i sędziów przed przemocą online



2.3. Sędziowanie: AI na linii spalonego

Podstawowym wsparciem dla arbitrów był system Advanced Semi-Automated Offside Technology. Na każdym stadionie działało 16 kamer śledzących zawodników i piłkę, rejestrujących do 29 punktów ciała każdego gracza 50 razy na sekundę. Dane z kamer łączono z odczytami czujnika inercyjnego umieszczonego w oficjalnej piłce TRIONDA, który z dużą precyzją określał moment zagrania.

W porównaniu z mundialem w Katarze system mógł automatycznie przekazywać sędziom informacje o jednoznacznych spalonych pozycyjnych, przyspieszając podejmowanie decyzji. Nie oznaczało to jednak pełnej automatyzacji. Ocena, czy zawodnik aktywnie uczestniczył w akcji, utrudniał grę przeciwnikowi lub zastępował bramkarza, nadal należała do sędziego i zespołu VAR.

2.4. Cyfrowe awatary zawodników

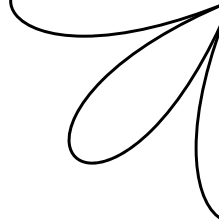
Przed turniejem wykonano trójwymiarowe skany wszystkich 1 248 zawodników. Skanowanie jednej osoby trwało około sekundy i pozwalało odwzorować wymiary oraz proporcje poszczególnych części ciała.

Modele 3D miały dwa zastosowania. Po pierwsze, dostarczały dodatkowych danych systemowi śledzenia, szczególnie gdy część sylwetki zawodnika była zasłonięta przez innego gracza. Po drugie, awatary wykorzystywano w animacjach prezentujących decyzje o spalonym kibicom na stadionie i widzom transmisji. Zamiast jednakowych, uproszczonych figur można było pokazywać modele przypominające rzeczywistych uczestników akcji.



2.5.

Generatywna AI w analizie taktycznej



Najważniejszym narzędziem dla reprezentacji był FIFA AI Pro, opracowany przez FIFA i Lenovo. System, udostępniony wszystkim 48 drużynom, łączył generatywną AI z danymi meczowymi, nagraniami wideo i informacjami o pozycjach zawodników. Analizował miliony punktów danych oraz ponad 2 tys. wskaźników, odpowiadając na pytania trenerów w formie raportów, wykresów, klipów wideo i wizualizacji 3D.

Narzędzie wspierało analizy przed- i pomeczowe, pozwalając szybko wyszukiwać informacje o taktyce przeciwnika i własnej grze. Dzięki temu wszystkie federacje otrzymały dostęp do zaawansowanej analityki opartej na oficjalnych danych turniejowych, niezależnie od wielkości sztabu i budżetu.

2.6.

Referee View: mecz oczami sędziego

Jedną z najbardziej widowiskowych nowości był system Referee View, czyli kamera umieszczona na głowie głównego arbitra. Algorytmy AI stabilizowały obraz w czasie rzeczywistym, ograniczając drgania i pozwalając pokazywać akcje z perspektywy sędziego.

Ujęcia wykorzystywano we wszystkich meczach fazy grupowej, średnio około trzy razy na spotkanie. Technologia nie tylko zwiększała atrakcyjność transmisji, ale także pomagała widzom zrozumieć warunki pracy arbitra i tempo podejmowania decyzji.



Referee View był używany we wszystkich meczach fazy grupowej. Według wstępnych danych FIFA ujęcia pojawiały się przeciętnie około trzech razy w jednym spotkaniu i były dostępne dla nadawców na ponad 220 terytoriach.

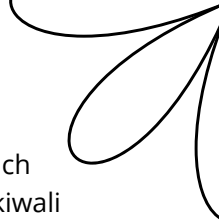
W obsługę technologiczną zaangażowano ponad 17 tys. urządzeń Lenovo i Motorola oraz przeszło 200 inżynierów. Nie wszystkie te rozwiązania były same w sobie sztuczną inteligencją, ale tworzyły infrastrukturę konieczną do działania modeli analitycznych, przetwarzania obrazu i szybkiej dystrybucji treści.

2.7. Operacje stadionowe: cyfrowe bliźniaki w akcji

Rozgrywanie turnieju w trzech państwach i 16 miastach wymagało jednoczesnego monitorowania stadionów, centrów treningowych, transportu, zaplecza technicznego i działań personelu. W tym celu w Tournament Operations Center w Miami działało Intelligent Command Center. System łączył dane operacyjne z różnych obszarów i prezentował je w jednym środowisku. Kierownictwo FIFA mogło otrzymywać alerty, obserwować powtarzające się problemy oraz koordynować reakcję wielu zespołów.

AI przygotowywała również codzienne podsumowania operacyjne i pomagała identyfikować trendy występujące w różnych lokalizacjach. Uzupełnieniem systemu były cyfrowe bliźniaki stadionów i innych obiektów turniejowych, czyli ich wirtualne modele zasilane danymi operacyjnymi. Pozwalały one wizualizować sytuację na obiekcie, analizować przepływ kibiców, identyfikować potencjalne zatory komunikacyjne oraz symulować skutki różnych scenariuszy, np. zmian tras dojścia czy zamknięcia wybranych stref. Dzięki temu organizatorzy mogli szybciej reagować na problemy i lepiej planować działania służb porządkowych, transportowych oraz obsługi stadionów.





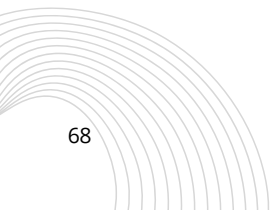
Cyfrowe bliźniaki współpracowały również z narzędziami nawigacyjnymi wykorzystującymi informacje o stadionach, strefach kibica i punktach zainteresowania. W efekcie organizatorzy uzyskiwali wspólny obraz sytuacji w 16 miastach-gospodarzach, a kibice mogli sprawniej poruszać się po obiektach.

Należy jednak podkreślić, że nie był to autonomiczny system zarządzający turniejem. AI wspierała agregowanie danych, wykrywanie anomalii i przygotowywanie rekomendacji, natomiast decyzje podejmowali pracownicy FIFA oraz lokalne zespoły operacyjne.

2.8. | Bezpieczeństwo cyfrowe: AI przeciwko przemocy online

Jednym z największych ilościowo zastosowań AI podczas mundialu był FIFA Social Media Protection Service. System monitorował treści kierowane do zawodników, trenerów, sędziów i oficjalnych profili drużyn. Do 18 lipca usługa moderowała ponad 53 mln wpisów i komentarzy. Zidentyfikowano ponad 7 mln potencjalnie szkodliwych lub obraźliwych treści, a ponad 200 tys. materiałów zgłoszono lub objęto dodatkowymi działaniami. Ponad pół miliona komunikatów wykrytych przez AI jako potencjalnie niebezpieczne zostało następnie ocenionych przez zespół ekspertów. Ponad 15 tys. przypadków skierowano do dalszych działań, a przeszło tysiąc najpoważniejszych gróźb przekazano odpowiednim organom, w tym służbom porządkowym.

Jest to przykład właściwego modelu „human in the loop”: algorytm wykonuje masową selekcję treści, ale poważne decyzje, zwłaszcza dotyczące gróźb i zawiadamiania organów ścigania, wymagają oceny człowieka.



2.9. Korzyści z zastosowania AI

Rozwiązania AI wykorzystane podczas turnieju przyniosły korzyści zarówno na boisku, jak i poza nim. Umożliwiły szybsze podejmowanie decyzji sędziowskich, poprawiły zrozumienie gry przez kibiców, zapewniły reprezentacjom szerszy dostęp do narzędzi analitycznych oraz zwiększyły atrakcyjność transmisji. Równocześnie wspierały zarządzanie złożonymi operacjami organizacyjnymi i wzmacniały ochronę uczestników przed zagrożeniami w mediach społecznościowych.

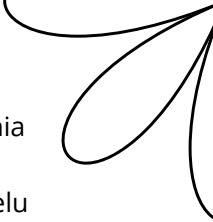
2.10. Ryzyka i kontrowersje

Rozwiązania AI wykorzystane podczas turnieju przyniosły korzyści zarówno na boisku, jak i poza nim. Umożliwiły szybsze podejmowanie decyzji sędziowskich, poprawiły zrozumienie gry przez kibiców, zapewniły reprezentacjom szerszy dostęp do narzędzi analitycznych oraz zwiększyły atrakcyjność transmisji. Równocześnie wspierały zarządzanie złożonymi operacjami organizacyjnymi i wzmacniały ochronę uczestników przed zagrożeniami w mediach społecznościowych.

2.10.1. Nadmierne zaufanie do algorytmów

Jednym z najczęściej wskazywanych zagrożeń jest tzw. automatyzacyjne uprzedzenie (automation bias), czyli skłonność ludzi do bezkrytycznego ufania rekomendacjom systemów informatycznych.





W przypadku technologii wykrywania spalonego szybkość działania i wysoka precyzja wizualizacji mogą sprawiać wrażenie całkowitej nieomyślności. W rzeczywistości poprawność wyniku zależy od wielu czynników, takich jak jakość obrazu z kamer, prawidłowa kalibracja systemu, synchronizacja danych czy poprawna interpretacja przepisów gry. Nawet najbardziej zaawansowane narzędzie powinno być traktowane jako wsparcie procesu decyzyjnego, a nie jego całkowite zastąpienie.

2.10.2. Brak pełnej przejrzystości modeli

FIFA i jej partnerzy technologiczni szczegółowo opisują funkcje poszczególnych rozwiązań, jednak nie publikują pełnej dokumentacji dotyczącej wykorzystywanych algorytmów. Nie są publicznie dostępne m.in. szczegółowe informacje na temat danych treningowych, procedur walidacji modeli, wskaźników błędów czy sposobów ograniczania potencjalnych uprzedzeń algorytmicznych. Utrudnia to niezależną ocenę skuteczności systemów oraz może rodzić pytania o transparentność procesów wpływających na przebieg zawodów sportowych.

2.10.3. Dane biometryczne i wizerunkowe

Tworzenie awatarów 3D wymaga bardzo dokładnego odwzorowania sylwetki zawodników. Powstające w ten sposób modele stanowią formę danych biometrycznych, które mogą posiadać wartość komercyjną także po zakończeniu turnieju. Rodzi to pytania o zakres zgody udzielanej przez sportowców, okres przechowywania danych, możliwość ich wykorzystania w przyszłych wydarzeniach, materiałach promocyjnych lub systemach treningowych, a także o prawa zawodników do ich cyfrowych reprezentacji. Wraz z rozwojem technologii generatywnej AI znaczenie tych kwestii będzie prawdopodobnie rosło.



2.10.4. Cyberbezpieczeństwo

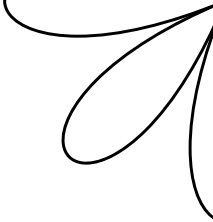
Współczesny turniej piłkarski opiera się na rozbudowanej infrastrukturze cyfrowej obejmującej systemy VAR, centra operacyjne, platformy analityczne, serwery transmisyjne oraz narzędzia komunikacyjne. Awaria techniczna lub skuteczny cyberatak mogłyby wpłynąć nie tylko na przebieg meczu, lecz także na transmisję telewizyjną, logistykę wydarzenia czy dostęp drużyn do danych analitycznych. Im większa jest zależność organizatorów od systemów cyfrowych, tym większe znaczenie zyskuje odporność infrastruktury na błędy, awarie i działania cyberprzestępców.

2.10.5. Ujednolicenie sposobu analizy

Jednym z mniej oczywistych ryzyk jest potencjalne ujednolicanie procesu analitycznego. Jeśli wszystkie reprezentacje korzystają z tego samego narzędzia, otrzymują podobne zestawy danych, wskaźników i rekomendacji. Może to prowadzić do częstszego koncentrowania się na tych samych elementach gry oraz poszukiwania podobnych wzorców taktycznych. Ostateczną przewagę nadal będą jednak osiągać sztaby potrafiące łączyć wyniki analiz AI z doświadczeniem trenerów, wiedzą o przeciwniku oraz własną interpretacją danych.

2.10.6. Błędy generatywnej AI

Systemy oparte na generatywnej sztucznej inteligencji mogą tworzyć odpowiedzi przekonujące, ale nie zawsze w pełni poprawne lub kompletne. Problem ten, określany często mianem „halucynacji AI”, może prowadzić do błędnych interpretacji danych, pomijania istotnego kontekstu lub nadmiernego upraszczania złożonych sytuacji boiskowych. W praktyce oznacza to, że narzędzia takie jak FIFA AI Pro powinny pełnić rolę asystenta wspierającego pracę analityków, a nie zastępować ekspercką ocenę trenerów i specjalistów przygotowania taktycznego.



2.10.7. | Ryzyko rosnącej zależności od technologii

Im więcej procesów zostaje powierzonych systemom cyfrowym, tym trudniejsze staje się funkcjonowanie bez nich. Dotyczy to zarówno sędziowania, jak i analiz sportowych, transmisji czy zarządzania turniejem. W dłuższej perspektywie może pojawić się pytanie, czy organizatorzy i uczestnicy nie staną się zbyt silnie uzależnieni od technologii, której awaria mogłaby znacząco zakłócić przebieg zawodów. Rozwój AI zwiększa możliwości organizacyjne i sportowe, ale równocześnie podnosi poziom złożoności całego ekosystemu piłkarskiego.

2.10.8. | Kluczowy wniosek

Korzyści z wykorzystania AI podczas mundialu są znaczące, jednak nie eliminują potrzeby ludzkiego nadzoru. Najważniejszym wyzwaniem pozostaje znalezienie równowagi pomiędzy automatyzacją a odpowiedzialnością człowieka za podejmowane decyzje. W praktyce skuteczność tych rozwiązań zależy nie tylko od jakości algorytmów, lecz także od przejrzystości ich działania, bezpieczeństwa infrastruktury i kompetencji osób korzystających z generowanych rekomendacji.



2.11. Podsumowanie

Mistrzostwa Świata 2026 nie były turniejem prowadzonym przez sztuczną inteligencję. Były jednak pierwszym mundialem, podczas którego AI stała się elementem niemal całego wydarzenia – od wykrywania spalonego i analizy gry przeciwnika, przez produkcję obrazu telewizyjnego, po logistykę i ochronę zawodników w internecie.

Najbardziej znaczące okazały się trzy kierunki:

1. **integracja danych z wielu źródeł** – kamer, piłki, materiałów wideo i baz statystycznych;
2. **udostępnienie generatywnej AI wszystkim zespołom** poprzez FIFA AI Pro;
3. **utrzymanie człowieka jako ostatecznego decydenta** w sędziowaniu, analizie i moderacji.

Mundial pokazał, że główną rolą AI w futbolu nie jest zastępowanie sędziów czy trenerów, lecz przetwarzanie ilości informacji przekraczającej możliwości człowieka. Przewaga będzie jednak zależała nie tylko od dostępu do algorytmów, ale od jakości danych, umiejętności zadawania właściwych pytań i zdolności krytycznej interpretacji odpowiedzi.



FIFA AI Pro – inteligentna platforma analityczna dla reprezentacji

FIFA AI Pro nie jest pojedynczym chatbotem, lecz wielowarstwową platformą analityczną łączącą generatywną AI, system wieloagentowy, analizę danych meczowych, przetwarzanie wideo oraz specjalistyczny model pojęciowy futbolu określany jako FIFA Football Language. Rozwiązanie, opracowane przez FIFA i Lenovo, zostało udostępnione wszystkim 48 reprezentacjom uczestniczącym w mundialu 2026.

System integrował trzy główne źródła danych: dane zdarzeniowe (np. podania, strzały, odbiory), tracking pozycyjny zawodników i piłki oraz oficjalne materiały wideo. Dzięki temu mógł analizować zarówno przebieg akcji, jak i ustawienie wszystkich uczestników gry w każdej fazie meczu. Kluczowym elementem systemu był FIFA Football Language, czyli ujednolicony system definiowania i klasyfikowania zdarzeń oraz faz gry w piłce nożnej. Dzięki niemu AI potrafiła powiązać pytania zadawane przez trenerów, takie jak „wysoki pressing” czy „progresja piłki”, z konkretnymi wskaźnikami, zdarzeniami i sytuacjami zapisanymi w bazach danych.

Według Lenovo system analizował miliony punktów danych oraz ponad 2 tys. wskaźników, wykorzystując architekturę wieloagentową. Poszczególne agenty odpowiadały za interpretację zapytania, planowanie analizy, wyszukiwanie danych, analizę taktyczną, dobór materiałów wideo i generowanie raportów. Dzięki temu użytkownik mógł zadawać pytania w języku naturalnym, a odpowiedzi otrzymywał w formie tekstu, wykresów, klipów meczowych, wizualizacji lub rekonstrukcji 3D.

FIFA AI Pro wykorzystywał mechanizm wyszukiwania wiedzy przypominający architekturę RAG (Retrieval-Augmented Generation). Oznacza to, że przed wygenerowaniem odpowiedzi system najpierw wyszukiwał odpowiednie dane i sytuacje meczowe, a dopiero później formułował wnioski. Rozwiązanie miało ograniczać ryzyko tzw. halucynacji AI i zapewniać osadzenie odpowiedzi w oficjalnych danych FIFA.

Technologicznie FIFA AI Pro stanowi przykład hybrydowej AI, łączącej bazy danych, algorytmy statystyczne i geometryczne, modele klasyfikacyjne, system wyszukiwania wiedzy, agentów AI, generatywne modele językowe oraz narzędzia wizualizacyjne. Generatywna AI odpowiadała przede wszystkim za interpretację pytań i prezentację wyników, natomiast właściwe obliczenia wykonywały wyspecjalizowane moduły analityczne.

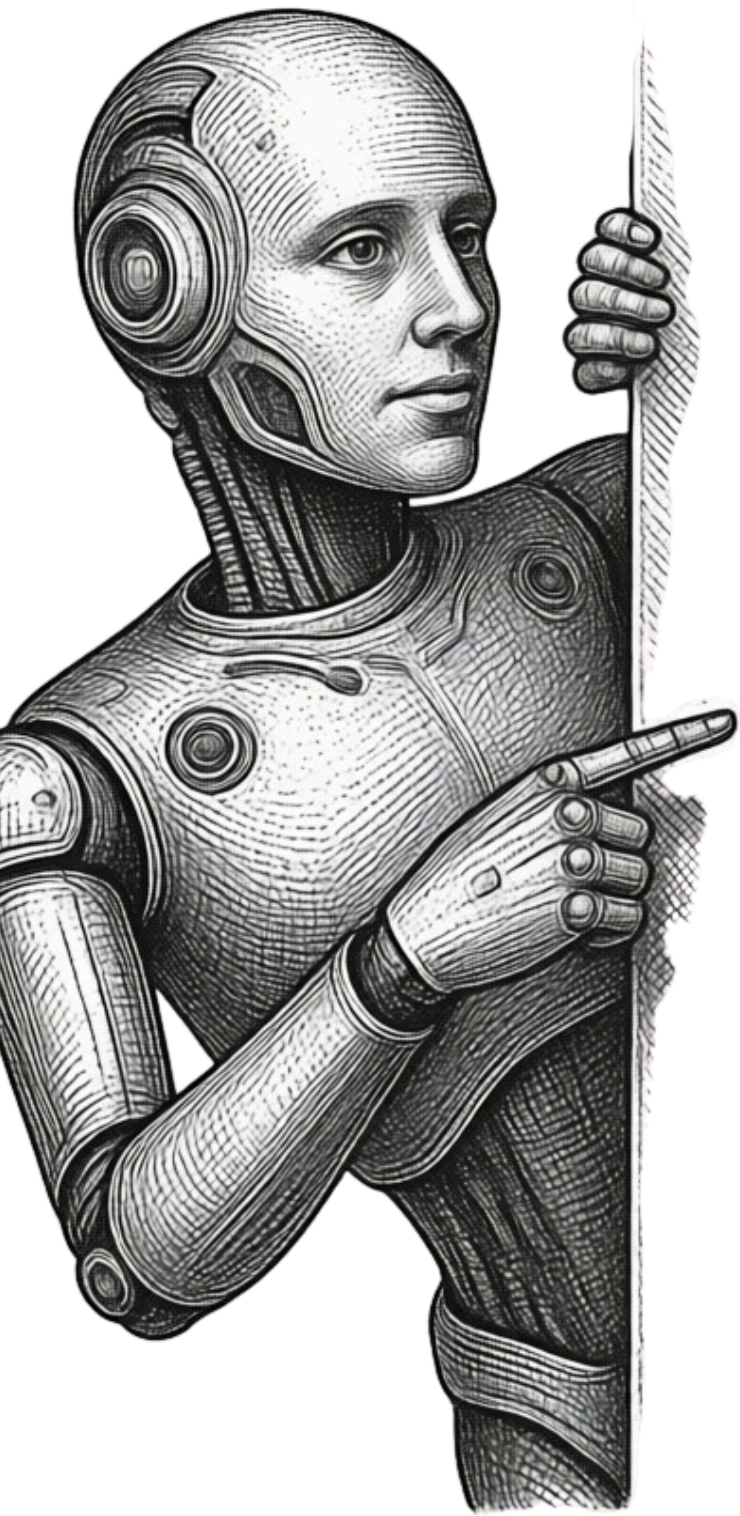
Najważniejszą innowacją FIFA AI Pro nie było więc zastosowanie samego modelu językowego, lecz stworzenie platformy integrującej standaryzowane dane meczowe, formalny język opisu futbolu, wieloagentową orkiestrację analiz oraz automatyczne wyszukiwanie materiałów wideo. System miał przyspieszać pracę sztabów szkoleniowych, ale nie zastępować analityków ani trenerów w podejmowaniu decyzji.



03

AI CZY SI?

**AKRONIM SZTUCZNEJ INTELIGENCJI W POLSKIEJ POLITYCE
SZTUCZNEJ INTELIGENCJI I PROGRAMIE OPERACYJNYM**



AKRONIM SZTUCZNEJ INTELIGENCJI W POLSKIEJ POLITYCE SZTUCZNEJ INTELIGENCJI I PROGRAMIE OPERACYJNYM

*ANALIZA PORÓWNAWCZA PRAKTYKI PAŃSTW CZŁONKOWSKICH UNII
EUROPEJSKIEJ WRAZ Z REKOMENDACJĄ TERMINOLOGICZNĄ*

3.1. Stan obecny — przypadek polskiej polityki krajowej

Polska polityka krajowa posługuje się w pełnej nazwie prawidłowym polskim terminem sztuczna inteligencja, lecz konsekwentnie skraca go do zapożyczonej z języka angielskiego formy AI. Nie jest to niedopatrzenie redakcyjne o charakterze incydentalnym. Rozbieżność została utrwalona w akcie prawnym.

Uchwała nr 196 Rady Ministrów z dnia 28 grudnia 2020 r. (M.P. 2021 poz. 23) w § 1 ustanawia „Politykę dla rozwoju sztucznej inteligencji w Polsce od roku 2020”, wprowadzając dla niej nazwę skróconą „Polityka AI”. Jest to definicja legalna nazwy skróconej, w której skrót nie ma oparcia w nazwie rozwiniętej: w polskim tytule dokumentu nie występuje ani litera „A”, ani litera „I”. Skrót nie skraca zatem niczego, ale zastępuje polską nazwę angielską.

Mechanizm tego zapożyczenia jest zresztą wprost ujawniony. Oficjalny portal sztucznej inteligencji administracji rządowej opisuje zakres Polityki jako działania „w obszarze sztucznej inteligencji (ang. Artificial Intelligence, «AI»)", po czym akronim wywiedziony z terminu angielskiego stosowany jest już wyłącznie w tekście polskim, w tym w nazwach sześciu obszarów interwencji: AI i społeczeństwo, AI i innowacyjne firmy, AI i sektor publiczny.

Praktyka ta budzi uzasadnione wątpliwości również z uwagi na tradycję piśmienniczą. W literaturze fachowej z zakresu informatyki, dziedziny inżynierskiej rozwijanej w Polsce nieprzerwanie od lat pięćdziesiątych XX wieku, akronimem przyjętym i utrwalałym jest SI.

3.2. Podstawa językowa

Słownik języka polskiego PWN definiuje sztuczną inteligencję jako „dział informatyki badający reguły rządzące zachowaniami umysłowymi człowieka i tworzący programy lub systemy komputerowe symulujące ludzkie myślenie”.

Źródło: <https://sjp.pwn.pl/sjp/sztuczna-inteligencja;2466532.html>

Skoro zatem obowiązującą w polszczyźnie nazwą dyscypliny jest sztuczna inteligencja, to konsekwentnym i jedynym poprawnym skrótem tej nazwy jest SI. Akronim AI stanowi skrót angielskiego wyrażenia *Artificial Intelligence* i jest właściwy dla dokumentów redagowanych w języku angielskim. Jego użycie w tekście polskim jest wewnętrznie niespójne: rozwinięcie skrótu nie odpowiada wówczas nazwie, którą skrót ma zastępować.

3.3. Forma fonetyczna - akronim, którego nie da się wymówić

Argument mający istotne znaczenie praktyczne, dotyczy wymowy. Dokument urzędowy nie funkcjonuje wyłącznie w postaci pisanej: jest odczytywany na posiedzeniach, cytowany w mediach, referowany w urzędach i omawiany w szkołach. Skrót musi mieć zatem ustaloną, powszechnie rozpoznawalną formę mówioną. Akronim AI takiej formy w polszczyźnie nie posiada.

Polskie literowce odczytuje się przy użyciu polskich nazw liter, jak: PKP jako pe-ka-pe, UE jako u-e, PZU jako pe-zet-u. Zastosowanie tej reguły do skrótu AI daje formę a-i, która jest homofoniczna ze spójnikiem „a i” i stanowi zbitkę dwóch samogłosek pozbawioną oparcia spółgłoskowego i w mowie ciągłej praktycznie nierozpoznawalną. Wzorec polski w tym przypadku zawodzi.

W konsekwencji mówiący sięgają po wzorec angielski, czyli ej aj, który wymaga kompetencji fonetycznej, jakiej znaczna część odbiorców nie posiada. Przy braku wprawy prowadzi to do form kuriozalnych, powszechnie dziś słyszanych w przestrzeni publicznej:

Spotykany wariant	Charakterystyka
„a-i”	Odczytanie zgodne z polskim wzorcem literowca. Homofoniczne ze spółnikiem „a i”; w mowie ciągłej nierozpoznawalne. Wzorzec formalnie poprawny, praktycznie nieużywalny.
„ej aj”	Odczytanie angielskie. Poprawne w języku źródłowym, lecz wymagające kompetencji fonetycznej, której znaczna część odbiorców nie posiada.
„aj”	Odczytanie zapisu jako dyftongu. Forma najczęstsza w mowie potocznej; całkowicie zrywa związek z nazwą.
„ej-i”, „a-aj”, „ej-a”	Formy pośrednie, powstałe z kontaminacji wzorca polskiego i angielskiego. Nieustabilizowane, wzajemnie niezrozumiałe.
„ejajowy”, „Al-owy”, „ejajowiec”	Formy pochodne. Odmieniane wbrew normie lub nieodmienne; brak akceptowalnego przymiotnika i nazwy wykonawcy czynności.

Skutek jest podwójny. Po pierwsze, skrót przestaje pełnić funkcję skrótu: słuchacz, który usłyszy „ej aj” lub „aj”, nie ma żadnej możliwości odtworzenia z tej formy nazwy „sztuczna inteligencja”. Więż między skrótem a rozwinięciem, wątpliwa już w piśmie, w mowie zanika całkowicie. Po drugie, brak normy wymawianiowej ma **wymiar społeczny**: osoba nieznająca angielskiego, próbując posłużyć się terminem z dokumentu adresowanego do niej samej, naraża się na ośmieszenie. Jest to bariera nie tylko komunikacyjna, lecz i godnościowa.

3.4. Praktyka porównawcza państw członkowskich UE

W wielu państwach Europy krajowe polityki sztucznej inteligencji publikowane są dwujęzycznie; część dokumentów dostępna jest wyłącznie w języku angielskim, część, wyłącznie w języku urzędowym danego państwa. Zestawienie zamieszczone w dalszej części dokumentu obejmuje polityki i strategie wdrażania sztucznej inteligencji opracowane przez rządy państw członkowskich Unii Europejskiej oraz oficjalne publikacje instytucji rządowych. Analiza tych dokumentów pozwala wyróżnić trzy modele postępowania.

Model I — dokument w języku urzędowym, akronim rodzimy. Model dominujący. Państwo redaguje dokument we własnym języku i posługuje się skrótem wywiedzionym z własnej nazwy, bądź konsekwentnie stosuje pełną nazwę bez skrótu. Należą tu m.in. Austria i Niemcy (KI), Francja, Hiszpania, Włochy, Rumunia i Portugalia (IA), Słowenia i Chorwacja (UI), Węgry i Łotwa (MI), Litwa (DI), Bułgaria (II), a także Dania, Estonia i Finlandia, gdzie akronimu nie stosuje się wcale.

Model II — dokument w języku angielskim, akronim AI. Model w pełni uzasadniony. Jeżeli państwo publikuje politykę wyłącznie w języku angielskim, z uwagi na adresata międzynarodowego bądź dlatego, że angielski jest jego językiem urzędowym, stosowanie skrótu AI jest naturalne i poprawne, ponieważ odpowiada nazwie użytej w tym samym tekście. Dotyczy to Cypru, Grecji, Luksemburga, Irlandii i Malty. Model ten nie stanowi argumentu przeciwko akronimowi rodzimemu, przeciwnie, potwierdza regułę spójności nazwy i skrótu.

Model III — dokument w języku urzędowym, akronim angielski. Model najmniej liczny i jedyny wewnętrznie niespójny. Państwo redaguje dokument we własnym języku, posługuje się pełną nazwą rodzimą, a następnie skraca ją do formy zapożyczonej, która z tej nazwy nie wynika. Do grupy tej należą Czechy i Polska.

Polska nie jest zatem przypadkiem odosobnionym, lecz należy do najmniej licznej grupy — tej, której praktyka nie znajduje uzasadnienia w wersji językowej dokumentu. Wybór akronimu AI byłby bowiem uzasadniony wyłącznie wtedy, gdyby Polityka została opublikowana wyłącznie po angielsku.

Osobnego wyjaśnienia wymagają Szwecja, Holandia, Irlandia i Malta. Występujący tam akronim AI nie jest zapożyczeniem: w Szwecji wynika wprost z nazwy szwedzkiej (artificiell intelligens), w Holandii z niderlandzkiej (artificiële intelligentie), a w Irlandii i na Malcie język angielski ma status języka urzędowego. Państwa te, podobnie jak pozostałe, stosują akronim spójny z nazwą. Potwierdza to regułę, zamiast ją podważać.

3.5. Zestawienie: akronimy sztucznej inteligencji w krajowych dokumentach polityki UE

TABELA OPRACOWANA NA PODSTAWIE KRAJOWYCH DOKUMENTÓW STRATEGICZNYCH ORAZ OFICJALNYCH PUBLIKACJI INSTYTUCJI RZĄDOWYCH PAŃSTW CZŁONKOWSKICH.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
1	Austria	niemiecki	Künstliche Intelligenz	KI	I	Artificial Intelligence Mission Austria 2030 (AIM AT 2030); KI-Umsetzungsplan 2024	W wersji niemieckojęzycznej konsekwentnie KI; akronim AI występuje wyłącznie w wersji angielskiej.
2	Belgia	niderlandzki, francuski, niemiecki	Artificiële Intelligentie / Intelligence artificielle / Künstliche Intelligenz	AI / IA / KI	I	AI 4 Belgium; National Convergence Plan for AI (2022)	Akronim dobrane do wersji językowej dokumentu.
3	Bułgaria	bułgarski	Изкуствен интелект	ИИ	I	Концепция за развитието на изкуствения интелект в България до 2030 г.	Akronim w języku urzędowym, zapisany cyrylicą.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
4	Chorwacja	chorwacki	Umjetna inteligencija	UI	—	brak przyjętego dokumentu krajowego	Na stronach rządowych spotykany również angielski akronim AI.
5	Cypr	grecki, turecki	Τεχνητή νοημοσύνη / Yapay zekâ	TN / YZ	II	National Artificial Intelligence Strategy (2020)	Dokument opublikowany w języku angielskim.
6	Czechy	czeski	Umělá inteligence	AI	III	Národní strategie umělé inteligence v ČR (NAIS)	Dokument w języku urzędowym, lecz posługujący się akronimem angielskim AI lub pełną nazwą czeską.
7	Dania	duński	Kunstig intelligens	KI	I	National strategi for kunstig intelligens	W dokumencie stosowana wyłącznie pełna nazwa w języku urzędowym.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
8	Estonia	estoński	Tehisintellekt	—	I	Eesti riiklik tehisintellekti alane tegevuskava („Kratikava”)	Stosowana pełna nazwa; akronim nieużywany.
9	Finlandia	fiński, szwedzki	Tekoäly / Artificiell intelligens	—	I	Tekoälyaiika; Artificial Intelligence 4.0	Stosowane pełne nazwy w obu językach urzędowych.
10	Francja	francuski	Intelligence artificielle	IA	I	Stratégie nationale pour l'Intelligence artificielle (France 2030)	Konsekwentnie akronim w języku urzędowym.
11	Grecja	grecki	Τεχνητή νοημοσύνη	TN	II	Blueprint for Greece's AI Transformation / Εθνική Στρατηγική για την TN	Wersja grecka posługuje się akronimem TN.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
12	Hiszpania	hiszpański	Inteligencia artificial	IA	I	Estrategia Nacional de Inteligencia Artificial (ENIA); Estrategia de IA 2024	Akronim w języku urzędowym, także w tytule dokumentu.
13	Holandia	niderlandzki	Kunstmatige intelligentie / Artificiële intelligentie	KI / AI	I	Strategisch Actieplan voor Artificiële Intelligentie (SAPAI); Overheidsbrede visie generatieve AI	Nazwa w języku urzędowym; akronim AI wywodzi się z niderlandzkiego „Artificiële intelligentie”.
14	Irlandia	irlandzki, angielski	Artificial intelligence	AI	II	AI – Here for Good: National AI Strategy (odświeżenie 2024)	Dokument w języku angielskim, będącym językiem urzędowym, akronim AI jest tu rodzimy.
15	Litwa	litewski	Dirbtinis intelektas	DI	I	Lietuvos dirbtinio intelekto strategija	Akronim w języku urzędowym.

Lp.	Państwo	Język(i) urzędowy(e)	"Sztuczna inteligencja" w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
16	Luksemburg	luksemburski, francuski, niemiecki	Kënschtlech Intelligenz / Intelligence artificielle / Künstliche Intelligenz	KI / IA	II	Artificial Intelligence: a strategic vision for Luxembourg	Dokument opublikowany w języku angielskim.
17	Łotwa	łotewski	Mākslīgais intelekts	MI	I	Informatīvais ziņojums „Par mākslīgā intelekta risinājumu attīstību”	Akronim w języku urzędowym.
18	Malta	maltański, angielski	Intelligenza artifizjali / Artificial intelligence	IA / AI	II	Malta: The Ultimate AI Launchpad – National AI Strategy 2030	Dokument w języku angielskim, będącym językiem urzędowym – akronim AI jest tu rodzimy.
19	Niemcy	niemiecki	Künstliche Intelligenz	KI	I	Strategie Künstliche Intelligenz der Bundesregierung (2018, aktualizacja 2020); KI-Aktionsplan	Konsekwentnie akronim w języku urzędowym.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
20	Polska	polski	Sztuczna inteligencja	AI	III	Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 (uchwała nr 196 RM z 28.12.2020, M.P. 2021 poz. 23)	Dokument w języku urzędowym pełna nazwa w języku polskim, akronim zapożyczony z angielskiego.
21	Portugalia	portugalski	Inteligência artificial	IA	I	AI Portugal 2030 – Estratégia Nacional para a Inteligência Artificial	Nazwa krajowa IA; w wersji angielskiej AI.
22	Rumunia	rumuński	Inteligența artificială	IA	I	Strategia Națională în domeniul Inteligenței Artificiale 2024-2027	Akronim w języku urzędowym.
23	Słowenia	słoweński	Umetna inteligenca	UI	I	Nacionalni program za umetno inteligenco (NpUI) do 2025	Akronim w języku urzędowym.

Lp.	Państwo	Język(i) urzędowy(e)	„Sztuczna inteligencja” w języku urzędowym	Akronim krajowy	Model	Krajowy dokument polityki dot. SI	Uwagi
24	Szwecja	szwedzki	Artificiell intelligens	AI	I	Nationell inriktning för artificiell intelligens; raport AI-kommissionen (2023)	Akronim AI jest zbieżny z angielskim, lecz wynika bezpośrednio z nazwy szwedzkiej.
25	Węgry	węgierski	Mesterséges intelligencia	MI	I	Magyarország Mesterséges Intelligencia Stratégiaja 2020–2030	Akronim w języku urzędowym.
26	Włochy	włoski	Intelligenza artificiale	IA	I	Strategia Italiana per l'Intelligenza Artificiale 2024–2026	Akronim w języku urzędowym.

Legenda kolumny „Model”: I — dokument w języku urzędowym, akronim rodzimy (lub wyłącznie pełna nazwa). II — dokument w języku angielskim, akronim AI (spójny z nazwą użytą w dokumencie). III — dokument w języku urzędowym, akronim angielski (nazwa i skrót niespójne). Wiersze wyróżnione kolorem oznaczają model III. „—” w kolumnie Model: brak przyjętego dokumentu krajowego.

3.6. Wnioski

Przedstawiona analiza prowadzi do czterech wniosków.

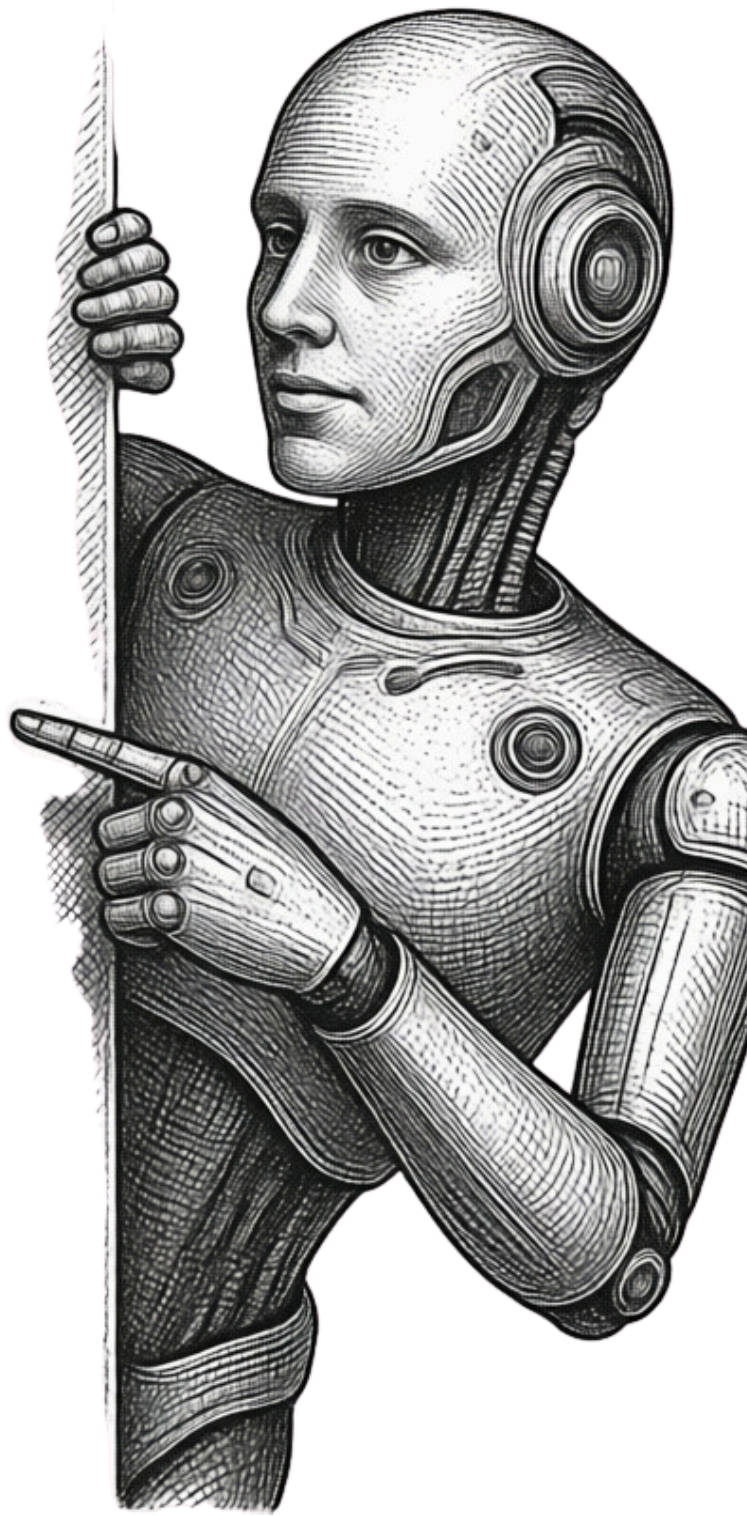
Po pierwsze — niespójność obowiązujących dokumentów. Polska polityka krajowa posługuje się pełną nazwą polską, lecz skraca ją do akronimu wywiedzionego z języka angielskiego. Skrót „AI” nie jest skrótem od „sztucznej inteligencji”, nie zawiera żadnej z liter tej nazwy. Uchwała nr 196 Rady Ministrów utrzymała tę rozbieżność w formie definicji legalnej nazwy skróconej.

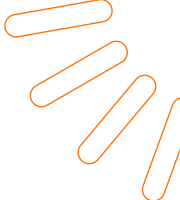
Po drugie — norma językowa. Polska nazwa dyscypliny została zdefiniowana słownikowo, a jej jedynym konsekwentnym skrótem jest SI.

Po trzecie — praktyka europejska. Dominującym modelem w Unii Europejskiej jest akronim spójny z językiem, w którym zredagowano dokument. Państwa publikujące politykę wyłącznie po angielsku posługują się skrótem AI zasadnie, Polska takiego uzasadnienia nie ma, w wersji Polityki opublikowanej po polsku. Przyjęcie akronimu SI nie byłoby zatem odstępstwem od standardu europejskiego, lecz powrotem do niego.

Po czwarte — zrozumiałość i wymawialność. Jest to argument o największym ciężarze. Dla znacznej części społeczeństwa, w szczególności dla starszych pokoleń Polaków, zapis „SI” jest bezpośrednio czytelny, ponieważ odsyła do znanej polskiej nazwy, i ma jednoznaczną formę mówioną („es-i”). Skrót „AI” nie ma w polszczyźnie ustalonej wymowy: wzorzec polski daje nierozpoznawalne „a-i”, wzorzec angielski wymaga kompetencji, której wielu odbiorców nie ma, a w praktyce funkcjonuje kilka wzajemnie niezrozumiałych wariantów. W dokumencie adresowanym do ogółu obywateli jest to bariera komunikacyjna, a nie kwestia estetyki językowej.

04

MIESIĘCZNE**PODSUMOWANIE TRENDÓW****TECHNOLOGICZNYCH (LIPIEC 2026)**



04 | MIESIĘCZNE PODSUMOWANIE TRENDÓW TECHNOLOGICZNYCH (LIPIEC 2026)

Analiza danych z raportów tygodniowych za lipiec 2026 roku wskazuje na wyraźne przyspieszenie w obszarach rozbudowy infrastruktury obliczeniowej, zaostreżania wymogów regulacyjnych w Unii Europejskiej oraz wdrażania kryptografii postkwantowej. Rywalizacja technologiczna przeniosła się z poziomu samych modeli algorytmicznych na poziom kontroli nad danymi, odporności łańcuchów dostaw półprzewodników i bezpiecznego środowiska wykonawczego.

4.1. | Sztuczna Inteligencja (AI): Interoperacyjność, Modele Dedykowane i Suwerenność Danych

- **Regulacje Interoperacyjności i Dostępu do Danych:** Komisja Europejska narzuciła Google wiążące środki w ramach aktu o rynkach cyfrowych (DMA).

Zobowiązują one do bezpłatnego udostępnienia konkurentom 11 kluczowych funkcji systemu Android (do wydań Android 18 i 19 w latach 2027–2028) oraz zanonimizowanych danych wyszukiwania do poprawy interpretacji zapytań. Równolegle KE otworzyła konsultacje w sprawie suwerenności danych, celem ograniczenia zależności od państw trzecich.

- **Nowa Generacja Modelowa i Specjalizacja:** OpenAI wprowadziło rodzinę modeli GPT-5.6 (warianty Sol, Terra, Luna), kładąc nacisk na programowanie, automatyczną naprawę kodu i cyberbezpieczeństwo.
- **Lokalne Ekosystemy i Suwerenne Dane:** W Polsce ruszyła społecznościowa zbiórka danych wizualnych poprzez aplikację „Obywatel Bielik” (Fundacja SpeakLeash), mająca dostarczyć 1 mln zdjęć osadzonych w polskich realiach do treningu modelu Bielik.AI.
- **Kryzysowa i Medyczna Aplikacja AI:** Rozwijane są systemy AI wspierające odporność na katastrofy naturalne (np. Google Flood Hub, wykraczający do 2 mld ludzi, oraz konstelacja FireSat). W medycynie precyzyjnej wykorzystuje się superkomputery (program NSF ACCESS, DeltaAI) do analizy wariantów genetycznych i prognozowania ekspresji genów.
- **Wykonywanie zadań w świecie fizycznym (Physical AI):** Postępuje integracja generatywnej AI z robotyką. Chiny narzucają globalne tempo, wdrażając kompleksowy system standardów dla robotów humanoidalnych.

Implikacje:

- **Rynek:** Wymuszona przez DMA interoperacyjność wielkich ekosystemów otwiera przestrzeń dla mniejszych dostawców usług AI, którzy dotychczas odbijali się od barier wejścia na platformy mobilne.
- **Wpływ Strategiczny:** Rośnie wartość zlokalizowanych, unikalnych zbiorów danych (jak projekt „Obywatel Bielik”). Z perspektywy administracji publicznej kluczowe staje się kategoryzowanie i udostępnianie krajowych repozytoriów danych w sposób kontrolowany, co jest niezbędnym fundamentem do trenowania suwerennych modeli algorytmicznych niezależnych od zewnętrznych dostawców.

4.2.

Infrastruktura Obliczeniowa, HPC i Gigafabryki AI

- **Polski Krok w Stronę Gigafabryki AI:** Rada Ministrów przyjęła uchwałę otwierającą Polsce drogę do udziału w europejskim programie Gigafabryk AI (EuroHPC JU). Deklaracja zakłada zakup usług AI o wartości ponad 400 mln zł w latach 2028–2033, co może przyciągnąć do kraju inwestycje w infrastrukturę o wartości przekraczającej 10 mld zł (z potencjałem do 75 tys. akceleratorów GPU).
- **Europejski Wyścig Infrastrukturalny:** W Unii Europejskiej realizowane są projekty wykonawcze, takie jak francuskie konsorcjum ÆTHER w Strasburgu oraz działania polskiego Euvic Solutions z Nvidią (budowa fabryki AI w Grecji w oparciu o architekturę SuperPOD).
- **Integracja HPC i Kwantów:** Prowadzone są międzynarodowe projekty łączące moc obliczeniową superkomputerów z systemami kwantowymi (np. europejsko-japoński projekt QNeko).

Implikacje:

- **Rynek:** Potencjalne uruchomienie inwestycji rzędu 10 mld zł w polską infrastrukturę AI (w ramach EuroHPC) stworzy potężny popyt na usługi inżynieryjne i budowlane w sektorze Data Center.
- **Wpływ Strategiczny:** Głównym wąskim gardłem dla tych projektów staje się podaż energii. Lokowanie obiektów o zapotrzebowaniu na moc rzędu dziesiątek megawatów wymaga precyzyjnych analiz sieci przesyłowej. Oceniając potencjał absorpcji tak dużych obiektów przez warszawski ekosystem innowacji, decydująca będzie modernizacja stołecznego ringu energetycznego oraz gwarancje dostaw zielonej energii. Bez tego zasilenie planowanych klastrów (np. 75 tys. akceleratorów GPU) na poziomie lokalnym będzie barierą nie do przejścia.



- **Skalowanie Pamięci dla AI:** Koreańscy naukowcy opracowali metodę tworzenia ultracienkich stosów pamięci trójwymiarowych (ponad 10 warstw chipów o grubości kilkunastu mikrometrów). Rozwiązanie to skraca drogę przepływu danych do układów obliczeniowych, stanowiąc odpowiedź na ograniczenia tradycyjnych architektur.
- **Łańcuchy Dostaw i Sprzęt Produkcyjny:** Micron zapowiedział inwestycję do 3 mld USD w USA (w tym 500 mln USD dla GlobalWafers na fabrykę wafla krzemowych 300 mm w Teksasie). Raporty rynkowe wskazują, że fundamentem branży pozostaje ok. 25 producentów urządzeń produkcyjnych (np. ASML, NAURA).
- **Pozycja Polski:** Polska umacnia swoją pozycję w europejskim ekosystemie półprzewodników dzięki nowym centrom B+R i zacieśnianiu współpracy z Tajwanem.
- **Przełom w Chłodzeniu:** Zidentyfikowano nowy związek tantalu i azotu o przewodności cieplnej przewyższającej miedź, co stwarza możliwości zastosowania go w charakterze warstw pośrednich w chłodzeniu gęsto upakowanej elektroniki.

Implikacje:

- **Rynek:** Wyścig o układanie pamięci w 3 wymiarach (3D) i nowe materiały chłodzące dla gęsto upakowanej elektroniki premiują firmy inwestujące w zaawansowane R&D. Skraca to cykl życia obecnych generacji sprzętu obliczeniowego.
- **Wpływ Strategiczny:** Zacieśnianie współpracy badawczej i gospodarczej stwarza szansę na przyciągnięcie do kraju zaawansowanych centrów projektowych SC. Lokalne ośrodki badawcze mają szansę wpiąć się w globalne łańcuchy dostaw jako huby testowe, wyspecjalizowane np. w termodynamice zaawansowanych układów skalonych.

4.4.

Cyberbezpieczeństwo i Kryptografia Postkwantowa (PQC)

- **Presja Regulacyjna w USA (PQC do 2030):** Rozporządzenie wykonawcze Białego Domu (EO 14412) wyznaczyło szczytny termin 2030 roku na pełne wdrożenie standardów kryptografii postkwantowej FIPS NIST w infrastrukturze federalnej i u jej dostawców. Decyzja ta wymusza pilną integrację sprzętowych akceleratorów PQC i architektur hybrydowych.
- **Działania Unii Europejskiej:** KE zaprezentowała EU Action Plan on Cybersecurity and AI, przewidujący m.in. stworzenie europejskiej zdolności ewaluacji modeli zaawansowanych przed ich wprowadzeniem na rynek oraz budowę bezpiecznych środowisk testowych z udziałem agencji ENISA.
- **Testowanie Odporności i Wektory Ataków:** Badacze zaprezentowali kompletny, wieloetapowy atak na kwantową sieć neuronową (QNN) z wykorzystaniem struktury MITRE ATLAS oraz kanałów boczných. Z kolei Instytut Informatyki Teoretycznej i Stosowanej PAN wykorzystuje kwantowe modele generatywne do oceny podatności stosowanych mechanizmów szyfrowania.

Implikacje:

- **Rynek:** Szczytny termin (2030 r.) wyznaczony w USA dla infrastruktury federalnej stworzy globalny, lukratywny rynek na usługi audytorskie PQC oraz dostawy sprzętowych akceleratorów kryptograficznych.
- **Wpływ Strategiczny:** Sektor publiczny oraz operatorzy infrastruktury krytycznej muszą niezwłocznie rozpocząć inwentaryzację wykorzystywanych algorytmów szyfrujących. Konieczna jest rewizja wytycznych bezpieczeństwa m.in. dla zamówień publicznych, aby zablokować wdrażanie systemów, które w niedalekiej przyszłości będą podatne na ataki z wykorzystaniem maszyn kwantowych.

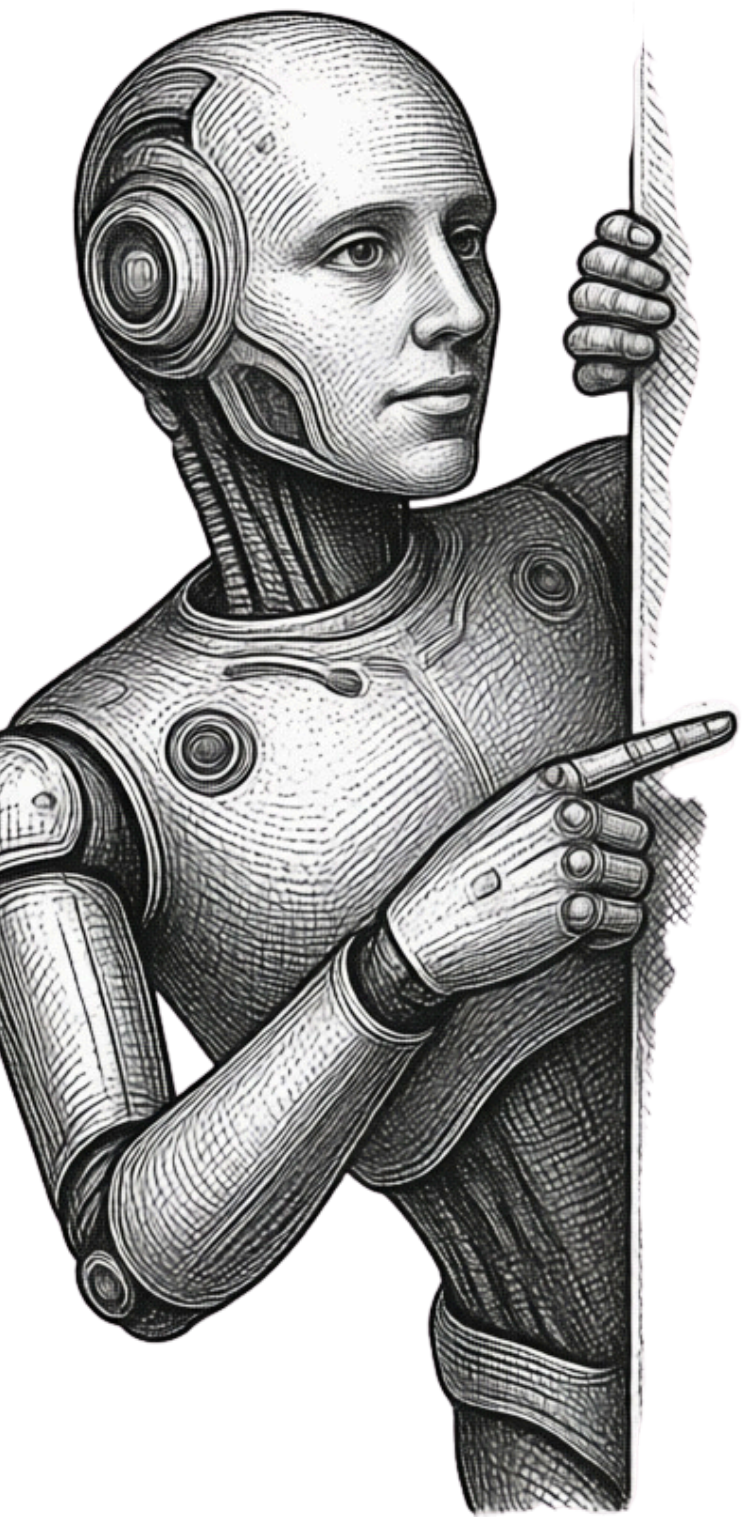


- **Zastrzeżenia KE wobec Mety (DSA):** Komisja Europejska wstępnie uznała, że mechanizmy zastosowane na Instagramie i Facebooku (m.in. nieskończone przewijanie, autoodtwarzanie, powiadomienia push) naruszają Akt o Usługach Cyfrowych (DSA) poprzez promowanie kompulsywnego użytkowania.
- **Unijna Standardyzacja Ochrony Dzieci:** Specjalny Panel KE ds. Bezpieczeństwa Dzieci w Internecie zalecił zharmonizowane ograniczenie dostępu do mediów społecznościowych dla dzieci poniżej 13 lat, wdrażanie bezpiecznych ustawień domyślnych oraz wykorzystanie dowodów z wiedzą zerową (ZKP) do weryfikacji wieku bez ujawniania tożsamości.
- **Krajowa Zależność Technologiczna:** Raport „Cyfrowa racja stanu” wykazał, że deficyt Polski w handlu produktami cyfrowymi przekroczył w 2025 r. 60 mld zł. Autorzy postulują wprowadzenie „Testu Suwerenności” przy zamówieniach publicznych, aby zapobiec zjawisku *vendor lock-in* i budować lokalne kompetencje.

Implikacje:

- **Rynek:** Rosnąca uwaga wokół deficytu w handlu produktami cyfrowymi (przekraczającego 60 mld zł) i dyskusja o „Teście Suwerenności” może drastycznie zmienić układ sił na rynku B2G (Business-to-Government), faworyzując europejskich dostawców kosztem globalnych hiperskalerów.
- **Wpływ Strategiczny:** Wdrożenie takich mechanizmów przy budowie państwowych systemów cyfrowych skutecznie ograniczyłoby zjawisko *vendor lock-in*. Długofalowo wymusiłoby to na zagranicznych korporacjach transfer technologii, a jednocześnie pozwoliłoby budować i utrzymywać wysoko wykwalifikowane kompetencje inżynierskie wewnątrz krajowego rynku.

05 | BIBLIOGRAFIA



- **Ewolucja cyberataków w erze AI:**

[1] D. Taler, "Reprompt: The Single-Click Microsoft Copilot Attack that Silently Steals Your Personal Data", Varonis Threat Labs, 16 czerwca 2026. [Online]. Dostępne: <https://www.varonis.com/blog/reprompt>. [dostęp: 28.07.2026].

[2] M. Eastland, "Anthropic accuses Alibaba of 'illicitly' accessing AI models", Bloomberg / The Japan Times, 25 czerwca 2026. [Online]. Dostępne: <https://www.japantimes.co.jp/business/2026/06/25/companies/anthropic-alibaba-illegal-access/>. [dostęp: 28.07.2026].

[3] "Agentic AI Security: \$4.7M Breaches, 92% Alarmed [2026]", Shattered Cyber Security Insights, 11 czerwca 2026. [Online]. Dostępne: <https://shattered.io/agentic-ai-security-2026/>. [dostęp: 28.07.2026].

[4] Phoenix Security, "Miasma Worm Reaches Microsoft Azure and PyPI: 73 Repositories Disabled, Hades Wave Drops 37 Malicious Python Wheels", Phoenix Security Supply Chain Intelligence, 5 czerwca 2026. [Online]. Dostępne: <https://phoenix.security/miasma-azure-hades-pypi-supply-chain-worm-2026/>. [dostęp: 28.07.2026].

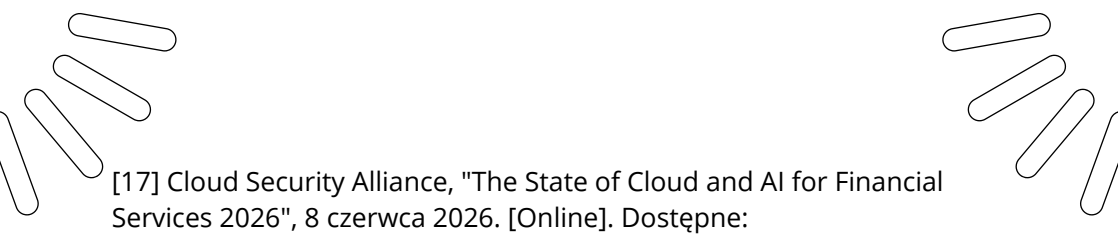
[5] Radware Cyberpedia, "Prompt Injection in 2026: Impact, Attack Types & Defenses", Radware Knowledge Hub, czerwiec 2026. [Online]. Dostępne: <https://www.radware.com/cyberpedia/prompt-injection/>. [dostęp: 28.07.2026].

[6] C. Schneider, "From LLM to agentic AI: prompt injection got worse", Christian Schneider Security Blog, 29 stycznia 2026 [Aktualizacja: czerwiec 2026]. [Online]. Dostępne: <https://christian-schneider.net/blog/prompt-injection-agentic-amplification/>. [dostęp: 28.07.2026].

[7] J. Martinez i L. Zhang, "Assessing Automated Prompt Injection Attacks in Agentic Environments", arXiv:2606.10525, 9 czerwca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2606.10525>. [dostęp: 28.07.2026].

[8] Modulos Governance Team, "OWASP Top 10 for Agentic Applications (2026) — Modulos Governance Guide", Modulos AI Documentation, 23 czerwca 2026. [Online]. Dostępne: <https://docs.modulos.ai/frameworks/owasp-top-10-agentic/>. [dostęp: 28.07.2026].

- [9] Futurum Group Labs, "Indirect Prompt Injection Exposes a Universal AI Security Flaw, No Deployment Model Is Immune", Futurum Research Insights, 9 czerwca 2026. [Online]. Dostępne: <https://futurumgroup.com/insights/indirect-prompt-injection-exposes-a-universal-ai-security-flaw-no-deployment-model-is-immune/>. [dostęp: 28.07.2026].
- [10] H. Gandhi, "Anthropic Accused Alibaba of a Distillation Attack. Here's What That Means—and Why It's So Dangerous", Inc. Magazine, 25 czerwca 2026. [Online]. Dostępne: <https://www.inc.com/hazel-gandhi/anthropic-accused-alibaba-of-a-distillation-attack-heres-what-that-means-and-why-its-so-dangerous/91365906>. [dostęp: 28.07.2026].
- [11] R. Nisha, "Inside Anthropic's Claims of Distillation Attack by Alibaba", Cyber Magazine, 25 czerwca 2026. [Online]. Dostępne: <https://cybermagazine.com/news/inside-anthropics-claims-of-distillation-attack-by-alibaba>. [dostęp: 28.07.2026].
- [12] "Anthropic accuses Alibaba of largest-ever attempt to extract Claude AI capabilities", Business Today, 25 czerwca 2026. [Online]. Dostępne: <https://www.businesstoday.in/technology/artificial-intelligence/story/anthropic-accuses-alibaba-of-largest-ever-attempt-to-extract-claude-ai-capabilities-539087-2026-06-25>. [dostęp: 28.07.2026].
- [13] Anthropic Research, "Detecting and preventing distillation attacks", Anthropic News, 23 lutego 2026 [Aktualizacja: czerwiec 2026]. [Online]. Dostępne: <https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks>. [dostęp: 28.07.2026].
- [14] WorkOS Blog, "AI agents now make up the majority of web traffic: What developers need to change", WorkOS, 18 czerwca 2026. [Online]. Dostępne: <https://workos.com/blog/ai-agent-web-traffic-what-developers-need-to-change>. [dostęp: 28.07.2026].
- [15] A. Choi, "AI Bots Surpass Human Traffic for First Time", The Chosun Daily, 24 czerwca 2026. [Online]. Dostępne: <https://www.chosun.com/english/industry-en/2026/06/24/UW7BO43H5BEX3HTRHY5VZEKAHY/>. [dostęp: 28.07.2026].
- [16] HUMAN Security, "The 2026 State of AI Traffic & Cyberthreat Benchmark Report", HUMAN Security Resources, czerwiec 2026. [Online]. Dostępne: <https://www.humansecurity.com/learn/resources/2026-state-of-ai-traffic-cyberthreat-benchmarks/>. [dostęp: 28.07.2026].



[17] Cloud Security Alliance, "The State of Cloud and AI for Financial Services 2026", 8 czerwca 2026. [Online]. Dostępne: <https://cloudsecurityalliance.org/artifacts/state-of-cloud-and-ai-for-financial-services-2026>. [dostęp: 28.07.2026].

[18] Group-IB Knowledge Hub, "Agentic AI Security in 2026: How Autonomous AI Agents Will Change Security Operations", Group-IB Reports, 23 czerwca 2026. [Online]. Dostępne: <https://www.group-ib.com/resources/knowledge-hub/agentic-ai-security/>. [dostęp: 28.07.2026].

[19] ACM Technology Policy Council, "Agentic AI: Autonomy, Opportunities, and Challenges of Action-Taking AI Systems", ACM TechBriefs, 11 czerwca 2026. [Online]. Dostępne: <https://www.acm.org/media-center/2026/june/techbrief-action-taking-ai>. [dostęp: 28.07.2026].

[20] Dimension Market Research, "Agentic AI Security Market", aktualizacja: kwiecień 2026. [Online]. Dostępne: <https://dimensionmarketresearch.com/report/agentic-ai-security-market/>. [dostęp: 28.07.2026].

[21] S. Everts i L. Scazzieri (red.), "Defending Europe, Deterring Russia: Resources, Readiness and Resolve", Chaillot Paper nr 190, European Union Institute for Security Studies (EUISS), czerwiec 2026. [Online]. Dostępne: https://www.iss.europa.eu/sites/default/files/2026-06/CP_190.pdf. [dostęp: 28.07.2026].

[22] B. Jensen, "Protecting Our Edge: Trade Secrets and the Global AI Arms Race", Center for Strategic and International Studies (CSIS), 7 maja 2025. [Online]. Dostępne: <https://www.csis.org/analysis/protecting-our-edge-trade-secrets-and-global-ai-arms-race>. [dostęp: 28.07.2026].

[23] Komisja Europejska, "AI Act: Regulatory framework for artificial intelligence", Shaping Europe's Digital Future, aktualizacja: 27 lipca 2026. [Online]. Dostępne: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>. [dostęp: 28.07.2026].

[24] E. Geller, "Anthropic warns state-linked actor abused its AI tool in sophisticated espionage campaign", Cybersecurity Dive, 28 stycznia 2026. [Online]. Dostępne: <https://www.cybersecuritydive.com/news/anthropic-state-actor-ai-tool-espionage/805550/>. [dostęp: 28.07.2026].

[25] Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign", Anthropic Security Report, 13 listopada 2025 [Cytowany w literaturze z 1 pol. 2026 r.]. [Online]. Dostępne:

<https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>. [dostęp: 28.07.2026].

[26] P. Hacker et al., "AI Agents Under EU Law: A Compliance Architecture for AI Providers", arXiv:2604.04604, 6 kwietnia 2026. [Online]. Dostępne: <https://arxiv.org/html/2604.04604v1>. [dostęp: 28.07.2026].

[27] TechPolicy.Press, "The EU AI Act is Not Ready for Agents", TechPolicy.Press, 5 maja 2026. [Online]. Dostępne: <https://www.techpolicy.press/the-eu-ai-act-is-not-ready-for-agents/>. [dostęp: 28.07.2026].

[28] Mishcon de Reya, "Agentic AI and cybersecurity: legal obligations and regulatory risks under UK and EU laws", Mishcon de Reya Legal Insights, 25 czerwca 2026. [Online]. Dostępne: <https://www.mishcon.com/news/agentic-ai-and-cybersecurity-legal-obligations-and-regulatory-risks-under-uk-and-eu-laws>. [dostęp: 28.07.2026].

[29] Steptoe LLP, "AI Agents: Navigating Emerging Risks and Compliance Challenges Under EU Law", Steptoe StepTechToe Blog, 23 czerwca 2026. [Online]. Dostępne: <https://www.steptoe.com/en/news-publications/steptechtoe-blog/ai-agents-navigating-emerging-risks-and-compliance-challenges-under-eu-law.html>. [dostęp: 28.07.2026].

[30] The White House, "Promoting Advanced Artificial Intelligence Innovation and Security" (Executive Order), Presidential Actions, 2 czerwca 2026. [Online]. Dostępne: <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>. [dostęp: 28.07.2026].

[31] S. C. Helms, J. D. Krieser, M. G. Morgan, A. H. Southwell i E. R. Golding, "New executive order shifts US AI policy toward national security", McDermott Will & Emery Insights, 9 czerwca 2026. [Online]. Dostępne: <https://www.mcdermottlaw.com/insights/new-executive-order-shifts-us-ai-policy-toward-national-security/>. [dostęp: 28.07.2026].

[32] M. McMurrough, C. Skeath i B. Ramirez, "CISA Releases Guidance on the Careful Adoption of Agentic AI Services", Global Policy Watch, 2 czerwca 2026. [Online]. Dostępne: <https://www.globalpolicywatch.com/2026/06/cisa-releases-guidance-on-the-careful-adoption-of-agentic-ai-services/>. [dostęp:

- [33] B. W. Setser, "How China Turned Its Platform Economy Into an AI Deployment Machine", Council on Foreign Relations, 1 czerwca 2026 [Aktualizacja: 22 czerwca 2026]. [Online]. Dostępne: <https://www.cfr.org/articles/chinas-platform-economy-is-an-ai-deployment-engine-the-u-s-is-still-looking-for-its-own>. [dostęp: 28.07.2026].
- [34] Baker McKenzie, "Singapore: IMDA updates Model AI Governance Framework for Agentic AI", Baker McKenzie Insight, czerwiec 2026. [Online]. Dostępne: <https://www.bakermckenzie.com/en/insight/publications/2026/06/singapore-imda-updates-model-ai-governance-framework-for-agentic-ai>. [dostęp: 28.07.2026].
- [35] "Singapore Updates Model AI Governance Framework for Agentic AI", Inside Global Tech, 18 czerwca 2026. [Online]. Dostępne: <https://www.insideglobaltech.com/2026/06/18/singapore-updates-model-ai-governance-framework-for-agentic-ai/>. [dostęp: 28.07.2026].
- [36] Information Commissioner's Office (ICO), "One year on: marking the 12-month commencement of the Data (Use and Access) Act", ICO News and Blogs, czerwiec 2026. [Online]. Dostępne: <https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2026/06/one-year-on-marking-the-12-month-commencement-of-the-data-use-and-access-act/>. [dostęp: 28.07.2026].
- [37] OpenAI, "OpenAI i Hugging Face wspólnie reagują na incydent bezpieczeństwa podczas ewaluacji modelu", OpenAI, 21 lipca 2026. [Online]. Dostępne: <https://openai.com/pl-PL/index/hugging-face-model-evaluation-security-incident/>. [dostęp: 28.07.2026].
- [38] Hugging Face Security Team, "Security incident – July 2026", Hugging Face Blog, 16 lipca 2026. [Online]. Dostępne: <https://huggingface.co/blog/security-incident-july-2026>. [dostęp: 28.07.2026].
- [39] Z. Wang et al., "ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?", arXiv:2605.11086, 11 maja 2026. [Online]. Dostępne: <https://arxiv.org/abs/2605.11086>. [dostęp: 28.07.2026].
- [40] R. Tiwari, S. Patil, S. Natarajan i A. Gowda, "Unpacking the AsyncAPI npm supply chain compromise and import-time payload delivery", Microsoft Security Blog, 15 lipca 2026. [Online]. Dostępne: <https://www.microsoft.com/en-us/security/blog/2026/07/15/unpacking-asyncapi-npm-supply-chain-compromise-import-time-payload-delivery/>. [dostęp: 28.07.2026].

[41] Unit 42, "The npm Threat Landscape: Attack Surface and Mitigations", Palo Alto Networks, aktualizacja 15 lipca 2026. [Online]. Dostępne: <https://unit42.paloaltonetworks.com/monitoring-npm-supply-chain-attacks/>. [dostęp: 28.07.2026].

[42] R. Prabhu, "Coordinated AsyncAPI Supply Chain Attack: Miasma RAT Delivered via Compromised CI/CD Pipelines in Two Repositories", StepSecurity, 14 lipca 2026. [Online]. Dostępne: <https://www.stepsecurity.io/blog/compromised-next-branch-pushes-malicious-asyncapi-generator-generator-helpers-and-generator-components-to-npm>. [dostęp: 28.07.2026].

[43] OWASP GenAI Security Project, "OWASP Top 10 for Agentic Applications for 2026", OWASP, 9 grudnia 2025. [Online]. Dostępne: <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>. [dostęp: 28.07.2026].

[44] Parlament Europejski i Rada Unii Europejskiej, "Dyrektywa (UE) 2022/2555 w sprawie środków na rzecz wysokiego wspólnego poziomu cyberbezpieczeństwa na terytorium Unii (NIS2)", Dz.U. L 333, 27 grudnia 2022, art. 21 ust. 2 lit. d. [Online]. Dostępne: <https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=CELEX:32022L2555>. [dostęp: 28.07.2026].

[45] C. Villa, A. E. Ozdarendeli, S. Tan i R. A. Popa, "Prismata: Confining Cross-Site Prompt Injection in Web Agents", arXiv:2607.08147, 9 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.08147>. [dostęp: 28.07.2026].

[46] Socket Research Team, "SANDWORM_MODE: Shai-Hulud-Style npm Worm Hijacks CI Workflows and Poisons AI Toolchains", Socket, 20 lutego 2026. [Online]. Dostępne: <https://socket.dev/blog/sandworm-mode-npm-worm-ai-toolchain-poisoning>. [dostęp: 28.07.2026].

[47] J. Prieto, "Denying the Worm: Detecting SANDWORM_MODE and the Emerging Class of AI Toolchain Supply Chain Attacks", CrowdStrike, 21 lipca 2026. [Online]. Dostępne: <https://www.crowdstrike.com/en-us/blog/denying-the-worm-sandworm-mode-and-ai-toolchain-supply-chain-attacks/>. [dostęp: 28.07.2026].

[48] D. Keller, "CrowdStrike Uncovers New Prompt Injection Techniques", CrowdStrike, 7 lipca 2026. [Online]. Dostępne: <https://www.crowdstrike.com/en-us/blog/crowdstrike-uncovers-new-prompt-injection-techniques/>. [dostęp: 28.07.2026].

- [49] A. Sasi, K. Dixit i A. K. Adimulam, "Indirect Prompt Injection in Web Content Targets AI Agents", Zscaler ThreatLabz, 2 lipca 2026. [Online]. Dostępne: <https://www.zscaler.com/blogs/security-research/indirect-prompt-injection-web-content-targets-ai-agents>. [dostęp: 28.07.2026].
- [50] M. Takahashi, "AgentForger, Part 1: ChatGPT Cross-Site Agent Forgery", Zenity Labs, 23 lipca 2026. [Online]. Dostępne: <https://labs.zenity.io/p/agentforger-part-1-chatgpt-cross-site-agent-forgery>. [dostęp: 28.07.2026].
- [51] M. Takahashi, "AgentForger, Part 2: The Autonomous Insider", Zenity Labs, 23 lipca 2026. [Online]. Dostępne: <https://labs.zenity.io/p/agentforger-part-2-the-autonomous-insider>. [dostęp: 28.07.2026].
- [52] F. Rosales, "When Your AI Reviewer Works for the Attacker: A Confused-Deputy Bug in Microsoft's Azure DevOps MCP Server", Manifold Security, 21 lipca 2026. [Online]. Dostępne: <https://www.manifold.security/blog/azure-devops-mcp-server-vulnerability>. [dostęp: 28.07.2026].
- [53] ESET Research, "ESET Threat Report: AI boosts cyber attackers' efficiency", ESET, 8 lipca 2026. [Online]. Dostępne: <https://www.eset.com/us/about/newsroom/research/eset-threat-report-ai-boosts-cyber-attackers-efficiency/>. [dostęp: 28.07.2026].
- [54] ESET Research, "ESET Threat Report H1 2026", ESET, lipiec 2026. [Online]. Dostępne: https://web-assets.eset.com/fileadmin/ESET/NL/Landingpage/2026/H1-2026_Threat-Report.pdf. [dostęp: 28.07.2026].
- [55] Z. Zhang, Z. Xie, W. Diao i J. Wu, "(A)I Sees What You Don't: Exploiting New Attack Surfaces in Third-Party Mobile Agents", arXiv:2607.00333, 1 lipca 2026 [Aktualizacja: 14 lipca 2026]. [Online]. Dostępne: <https://arxiv.org/abs/2607.00333>. [dostęp: 28.07.2026].
- [56] M. Dokhanian, "GhostApproval: A Trust Boundary Gap in AI Coding Assistants", Wiz Research, 8 lipca 2026. [Online]. Dostępne: <https://www.wiz.io/blog/ghostapproval-a-trust-boundary-gap-in-ai-coding-assistants>. [dostęp: 28.07.2026].
- [57] Hunt.io Research Team i B. Diachenko, "Thailand's Ministry of Finance Targeted With Hermes AI Agent Running Unattended, Hades Implant Staged", Hunt.io, 23 lipca 2026. [Online]. Dostępne: <https://hunt.io/blog/thailand-ministry-finance-targeted-with-hermes-ai-agent>. [dostęp: 28.07.2026].

- [58] G. Torres, S. Shrestha i S. Misra, "When Agents Remember Too Much: Memory Poisoning Attacks on Large Language Model Agents", arXiv:2607.06595, 6 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.06595>. [dostęp: 28.07.2026].
- [59] M. Clark, "JADEPUFFER: Agentic ransomware for automated database extortion", Sysdig Threat Research Team, 1 lipca 2026. [Online]. Dostępne: <https://www.sysdig.com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion>. [dostęp: 28.07.2026].
- [60] M. Clark, "JADEPUFFER evolves: The agentic threat actor deploys ransomware built to destroy AI models", Sysdig Threat Research Team, 20 lipca 2026. [Online]. Dostępne: <https://www.sysdig.com/blog/jadepuffer-evolves-the-agentic-threat-actor-deploys-ransomware-built-to-destroy-ai-models>. [dostęp: 28.07.2026].
- [61] B. Milanov i H. Khlaaf, "Friendly Fire: Hijacking Defensive Cyber AI Agents for Remote Code Execution", AI Now Institute, 8 lipca 2026. [Online]. Dostępne: <https://ainowinstitute.org/publications/friendly-fire-exploit-brief>. [dostęp: 28.07.2026].
- [62] W. Choi, J. Kim, T. Kang, J. Jeong, L. Xing i B. Lee, "Agent Data Injection Attacks are Realistic Threats to AI Agents", arXiv:2607.05120, 6 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.05120>. [dostęp: 28.07.2026].
- [63] B. Gyevnár, A. Kasirzadeh i N. B. Shah, "Distributed Denial of Science: How Indirect Data Poisoning of AI Systems Can Industrialize Scientific Fraud", arXiv:2607.10712, 12 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.10712>. [dostęp: 28.07.2026].
- [64] X. Zhou, P. Wang, Y. Zhou, W. Xie i X. Zhou, "Salience Induction against Multi-Hop RAG Agents: Threat and Defense", arXiv:2607.17535, 20 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.17535>. [dostęp: 28.07.2026].
- [65] K. Swinea, K. Aryal, L. Praharaj i M. Gupta, "VEXAIoT: Autonomous IoT Vulnerability EXploitation using AI Agents", arXiv:2607.09653, 10 lipca 2026. [Online]. Dostępne: <https://arxiv.org/abs/2607.09653>. [dostęp: 28.07.2026].
- [66] Boomi, "86% Of Enterprises Have Deployed AI Agents. Just 34% Trust Them, Boomi Study Finds", 20 lipca 2026. [Online]. Dostępne: <https://boomi.com/resources/resources-library/agentic-chaos-forrester-study/>. [dostęp: 28.07.2026].

[67] Cloud Security Alliance, "Defining Non-Human Identity", 22 lipca 2026. [Online]. Dostępne: <https://cloudsecurityalliance.org/artifacts/defining-non-human-identity>. [dostęp: 28.07.2026].

[68] R. Csernaton i P. Pawlak, "When AI Agents Attack: Autonomous Cyber Operations and Europe's Governance Gap", Carnegie Endowment for International Peace, 6 lipca 2026. [Online]. Dostępne: <https://carnegieendowment.org/research/2026/07/when-ai-agents-attack-autonomous-cyber-operations-and-europes-governance-gap>. [dostęp: 28.07.2026].

[69] Komisja Europejska, "EU Action Plan on Cybersecurity and Artificial Intelligence", 7 lipca 2026 [aktualizacja strony: 10 lipca 2026]. [Online]. Dostępne: <https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence>. [dostęp: 28.07.2026].

[70] Financial Times, "AI is becoming a geopolitical weapon, warns EU digital chief", 21 lipca 2026. [Online]. Dostępne: <https://www.ft.com/content/ccaf6c42-f94b-4afc-a063-7243b4fb78ed>. [dostęp: 28.07.2026].

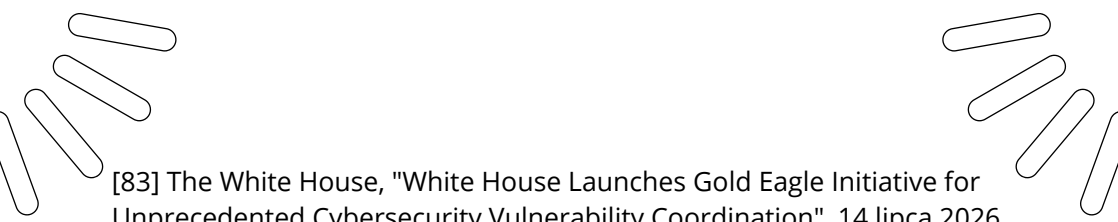
[71] Reuters, "China considers tighter export controls on AI models and chips", 21 lipca 2026. [Online]. Dostępne: <https://www.reuters.com/world/asia-pacific/china-considers-tighter-export-controls-ai-models-chips-ft-reports-2026-07-21/>. [dostęp: 28.07.2026].

[72] L. Chen, "As AI grows more powerful, a US-China feud threatens safety efforts", Reuters, 24 lipca 2026. [Online]. Dostępne: <https://www.reuters.com/legal/litigation/ai-grows-more-powerful-us-china-feud-threatens-safety-efforts-2026-07-24/>. [dostęp: 28.07.2026].

[73] E. Baptista, "China accuses US of 'AI hegemonism', threatens countermeasures over potential probes", Reuters, 27 lipca 2026. [Online]. Dostępne: <https://www.reuters.com/world/china/china-accuses-us-ai-hegemonism-threatens-countermeasures-over-potential-probes-2026-07-27/>. [dostęp: 28.07.2026].

[74] Parlament Europejski i Rada Unii Europejskiej, "Rozporządzenie (UE) 2026/1744 z dnia 8 lipca 2026 r. zmieniające rozporządzenia (UE) 2024/1689, (UE) 2018/1139 i (UE) 2023/1230 w odniesieniu do uproszczenia wdrażania zharmonizowanych przepisów dotyczących sztucznej inteligencji (Digital Omnibus on AI)", Dz.U. L, 24 lipca 2026. [Online]. Dostępne: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202601744. [dostęp: 28.07.2026].

- [75] Komisja Europejska, "Guidelines on transparency obligations for providers and deployers of AI systems", 20 lipca 2026. [Online]. Dostępne: <https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems>. [dostęp: 28.07.2026].
- [76] ENISA, "ENISA's view on Cybersecurity in the Frontier AI Era", 7 lipca 2026. [Online]. Dostępne: <https://www.enisa.europa.eu/publications/enisas-view-on-cybersecurity-in-the-frontier-ai-era>. [dostęp: 28.07.2026].
- [77] Komisja Europejska, "Commission publishes new guidance to support timely Cyber Resilience Act implementation", 27 lipca 2026. [Online]. Dostępne: <https://digital-strategy.ec.europa.eu/en/library/commission-publishes-new-guidance-support-timely-cyber-resilience-act-implementation>. [dostęp: 28.07.2026].
- [78] Parlament Europejski i Rada Unii Europejskiej, "Rozporządzenie (UE) 2024/2847 w sprawie horyzontalnych wymogów w zakresie cyberbezpieczeństwa w odniesieniu do produktów z elementami cyfrowymi (Cyber Resilience Act)", Dz.U. L, 20 listopada 2024. [Online]. Dostępne: https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=OJ:L_202402847. [dostęp: 28.07.2026].
- [79] Komisja Europejska, "Cyber Resilience Act - Reporting obligations", aktualizacja 8 czerwca 2026. [Online]. Dostępne: <https://digital-strategy.ec.europa.eu/en/policies/cra-reporting>. [dostęp: 28.07.2026].
- [80] Parlament Europejski, "Adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)", Legislative Observatory, status: projekt wycofany 6 października 2025. [Online]. Dostępne: <https://oeil.europarl.europa.eu/oeil/en/procedure-file?reference=2022%2F0303%28COD%29>. [dostęp: 28.07.2026].
- [81] Parlament Europejski i Rada Unii Europejskiej, "Dyrektywa (UE) 2024/2853 z dnia 23 października 2024 r. w sprawie odpowiedzialności za produkty wadliwe", Dz.U. L, 18 listopada 2024. [Online]. Dostępne: <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>. [dostęp: 28.07.2026].
- [82] UK Department for Science, Innovation and Technology oraz Lancaster University, "Thematic review and gap analysis on AI security", 10 lipca 2026. [Online]. Dostępne: <https://www.gov.uk/government/publications/thematic-review-and-gap-analysis-on-ai-security/thematic-review-and-gap-analysis-on-ai-security>. [dostęp: 28.07.2026].

- 
- [83] The White House, "White House Launches Gold Eagle Initiative for Unprecedented Cybersecurity Vulnerability Coordination", 14 lipca 2026. [Online]. Dostępne: <https://www.whitehouse.gov/releases/2026/07/white-house-launches-gold-eagle-initiative-for-unprecedented-cybersecurity-vulnerability-coordination/>. [dostęp: 28.07.2026].
- [84] Reuters, "White House monitors OpenAI's 'rogue' AI incident, lawmakers propose 'kill switch'", 23 lipca 2026. [Online]. Dostępne: <https://www.reuters.com/legal/litigation/ai-kill-switch-bill-floated-by-us-house-lawmakers-2026-07-23/>. [dostęp: 28.07.2026].
- [85] International Telecommunication Union, "ITU launches global standards initiative to build trust in AI behaviour and identity", 9 lipca 2026. [Online]. Dostępne: <https://www.itu.int/en/mediacentre/Pages/PR-2026-07-09-focus-group-agentic-AI.aspx>. [dostęp: 28.07.2026].
- [86] NVIDIA, "Industry Leaders Unite in Open Secure AI Alliance for AI Safety and Security", 27 lipca 2026. [Online]. Dostępne: <https://blogs.nvidia.com/blog/open-secure-ai-alliance/>. [dostęp: 28.07.2026].

- **Sztuczna inteligencja na Mistrzostwach Świata FIFA 2026:**

[1] FIFA, FIFA and Lenovo unveil multiple AI-powered innovations ahead of FIFA World Cup 2026, 7 stycznia 2026 r. – opis FIFA AI Pro, awatarów 3D zawodników oraz nowej generacji systemu Referee View.

[2] FIFA, Faster offside decisions, more stable referee body cams and more analysis: innovations at the FIFA World Cup 2026, 3 czerwca 2026 r. – podsumowanie najważniejszych technologii zastosowanych podczas turnieju.

[3] FIFA, Innovating the Game – oficjalny katalog technologii piłkarskich FIFA, obejmujący m.in. FIFA AI Pro, półautomatyczny spalony, śledzenie optyczne, połączoną piłkę, Referee View i rekonstrukcje 3D.

[4] FIFA, Semi-automated offside technology – opis sposobu działania półautomatycznego systemu wykrywania spalonego i roli zespołu VAR w podejmowaniu ostatecznej decyzji.

[5] FIFA, Connected ball technology – omówienie wykorzystania czujnika znajdującego się w piłce do precyzyjnego określenia momentu kontaktu zawodnika z piłką.

[6] FIFA, The FIFA World Cup 26 Official Match Ball: TRIONDA, 2 października 2025 r. – informacje o oficjalnej piłce mundialu i zastosowanym w niej czujniku ruchu działającym z częstotliwością 500 Hz.

[7] FIFA, 3D player scanning – opis procesu skanowania zawodników i tworzenia ich trójwymiarowych rekonstrukcji.

[8] FIFA, Lenovo technology enhances FIFA World Cup 2026 with referee cameras, player avatars and AI, 2026 r. – informacje o zeskanowaniu wszystkich 1 248 zawodników oraz wykorzystaniu awatarów w systemie półautomatycznego spalonego i wizualizacjach dla kibiców.

[9] FIFA, FIFA World Cup 2026 squads confirmed: 1,248 dreams, 48 teams and a truly global cast of players, 2 czerwca 2026 r. – oficjalne dane dotyczące liczby reprezentacji, zawodników i meczów turnieju.

[10] Lenovo, All 48 Teams Use FIFA AI Pro at FIFA World Cup 2026, 16 lipca 2026 r. – informacje o wykorzystaniu narzędzia przez wszystkie reprezentacje, analizowaniu milionów punktów danych i ponad 2 tys. wskaźników.



[11] Lenovo, Lenovo Technology Powers FIFA World Cup 2026 Operations and Strengthens AI-Driven Broadcast, 2 czerwca 2026 r. – opis FIFA AI Pro jako asystenta analitycznego dostępnego dla wszystkich 48 drużyn oraz infrastruktury wspierającej transmisje i operacje turniejowe.

[12] FIFA, FIFA and Lenovo unveil multiple AI-powered innovations ahead of FIFA World Cup 2026, 7 stycznia 2026 r. – charakterystyka Football AI Pro jako narzędzia łączącego generatywną AI, dane meczowe, materiały wideo i piłkarski model pojęciowy FIFA.

[13] FIFA, Referee View – oficjalny opis kamery noszonej przez sędziego i sposobu wykorzystania obrazu z perspektywy arbitra.

[14] Lenovo, A new perspective: how Lenovo AI is powering FIFA World Cup 2026 Referee View, lipiec 2026 r. – techniczny opis przetwarzania i stabilizowania obrazu z kamery sędziowskiej w czasie rzeczywistym.

[15] Lenovo, Opening goal of FIFA World Cup caught by Referee View, powered by Lenovo AI, lipiec 2026 r. – informacje o redukcji drgań obrazu oraz wykorzystaniu stabilizowanego sygnału przez producentów transmisji.

[16] Lenovo, FIFA World Cup Fans Embrace Referee View as Lenovo AI Brings Viewers Closer to the Pitch, 8 lipca 2026 r. – wyniki badania dotyczącego odbioru nowych technologii transmisyjnych przez widzów.

[17] Lenovo, Lenovo Technology Powers FIFA World Cup 2026 Operations and Strengthens AI-Driven Broadcast, 2 czerwca 2026 r. – informacje o infrastrukturze transmisyjnej, dystrybucji IPTV, urządzeniach i zapleczu technicznym turnieju.

[18] Lenovo, The Future of Football Is Here: AI Solutions to Power FIFA World Cup 2026, 6 stycznia 2026 r. – opis Inteligentnego Centrum Dowodzenia, monitoringu operacji oraz generowanych przez AI codziennych podsumowań sytuacyjnych.

[19] Lenovo, Breaking down the technology transforming the FIFA World Cup 2026, 22 czerwca 2026 r. – omówienie integracji danych dotyczących bezpieczeństwa, biletów, infrastruktury stadionowej, sprzedaży i obrazu z kamer.

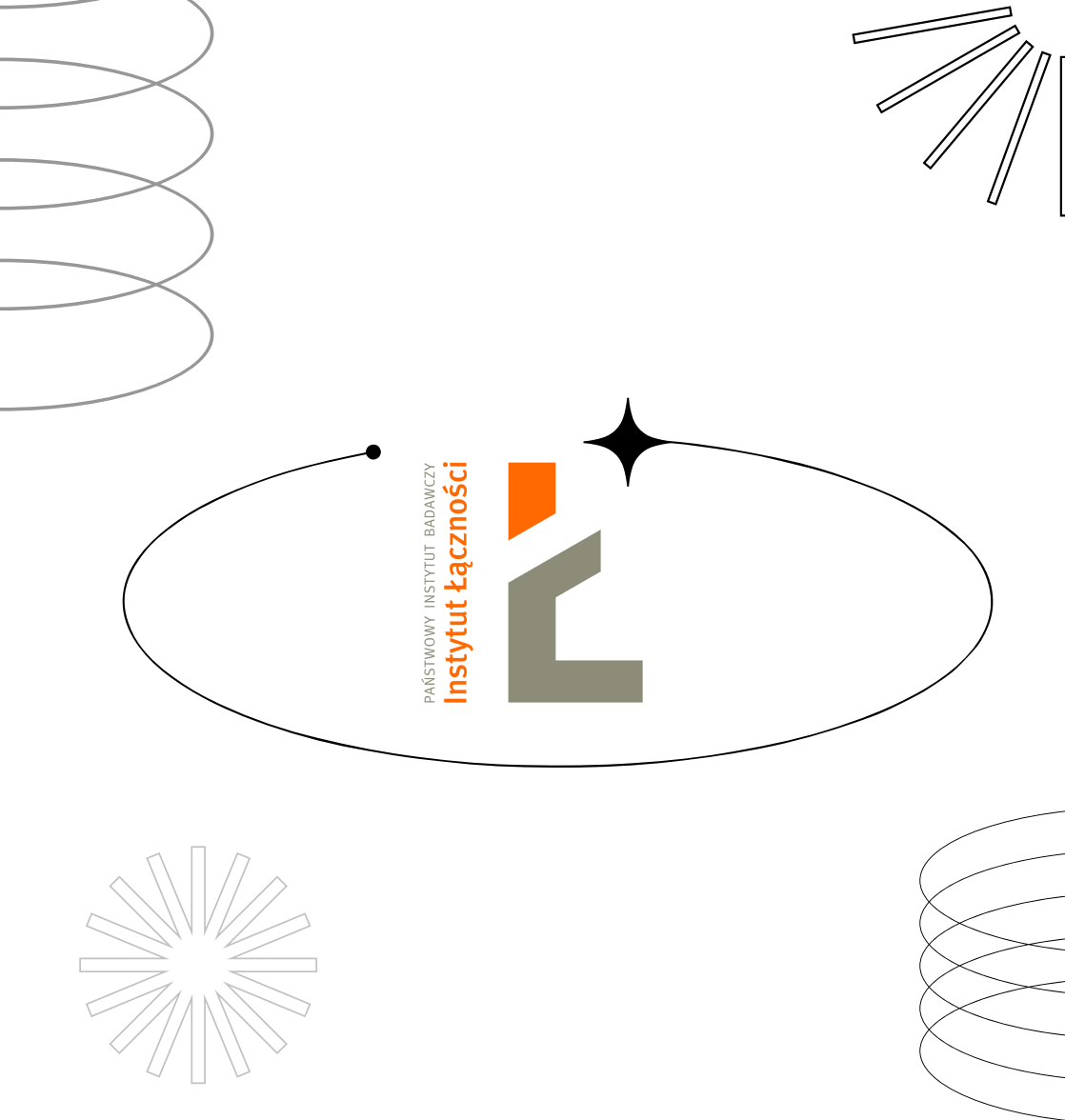


[20] Lenovo, AI at the speed of sport: real-time decisions and actions powered by Lenovo, 8 czerwca 2026 r. – syntetyczny opis systemów AI wspierających logistykę, analizę sportową, nawigację kibiców, transmisję i zarządzanie wydarzeniem.

[21] FIFA, FIFA moderates more than 53 million posts and comments during the FIFA World Cup 2026 through Social Media Protection Service, 18 lipca 2026 r. – końcowe dane dotyczące skali monitoringu treści, wykrywania nadużyć i przekazywania najpoważniejszych przypadków odpowiednim organom.

[22] FIFA, FIFA Social Media Protection Service identifies abusive posts during FIFA World Cup 2026 group stage, 1 lipca 2026 r. – dane częściowe z fazy grupowej oraz charakterystyka systemu wykrywania treści obraźliwych i dyskryminacyjnych.

[23] FIFA, FIFA releases new figures on Social Media Protection Service, 16 listopada 2024 r. – informacje o wcześniejszym rozwoju i wykorzystaniu systemu podczas turniejów FIFA, stanowiące kontekst dla wdrożenia z 2026 r.



PAŃSTWOWY INSTYTUT BADAWCZY
Instytut Łączności

Projekt finansowany ze środków Ministerstwa
Cyfryzacji. Publikacja wyraża jedynie poglądy
autora/ów i nie może być utożsamiana
z oficjalnym stanowiskiem Ministerstwa Cyfryzacji.



Projekt finansowany ze środków Ministerstwa Cyfryzacji

