CAROLINE WONG



AI w CYBERBEZPIECZEŃSTWIE

PRAKTYCZNY PRZEWODNIK PO ZAGROŻENIACH I STRATEGIACH OBRONNYCH ERY SZTUCZNEJ INTELIGENCJI

Helion

Tytuł oryginału: The AI Cybersecurity Handbook

Tłumaczenie: Piotr Pilch

ISBN: 978-83-289-4231-8

Copyright © 2026 by John Wiley & Sons, Inc.

All Rights Reserved. This translation published under license with the original publisher John Wiley & Sons, Inc.

Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Polish edition copyright © 2026 by Helion S.A.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiejkolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz wydawca dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz wydawca nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/aiwcyb
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

- Kup książkę
- Poleć książkę
- Oceń książkę

- Księgarnia internetowa
- Lubię to! » Nasza społeczność

Spis treści

O autorce 7

Przedmowa 9

Podziękowania 11

Wprowadzenie 13

CZĘŚĆ I. NOWY GRACZ W CYBERBEZPIECZEŃSTWIE: SZTUCZNA INTELIGENCJA PO OBU STRONACH

Rozdział 1. Era adaptacyjnych adwersarzy: wpływ sztucznej inteligencji na równowagę sił w bezpieczeństwie 23

Rozdział 2. Szybsze, tańsze i mądrzejsze rozpoznanie: czego hakerzy się dowiedzą, zanim Ty w ogóle to zauważysz 37

Rozdział 3. Od polimorfizmu do predykcyjności: jak sztuczna inteligencja wzmacnia złośliwe oprogramowanie z prędkością maszyn 57

Rozdział 4. Łamanie zasad: nowatorskie ataki i ograniczenia wykrywania sygnatur 83

Rozdział 5. Hiperrealistyczne i hiperukierunkowane: deepfake, phishing i oszustwa oparte na technologii AI 113

Kup książkę Poleć książkę

CZĘŚĆ II. BUDOWANIE ODPORNOŚCI ZA POMOCĄ SZTUCZNEJ INTELIGENCJI

Rozdział 6. Wzorce życia: uczenie się tego, co jest normalne, i wykrywanie tego, co takie nie jest 135

Rozdział 7. Od alertu do powstrzymywania — reagowanie na incydenty wspomagane przez technologię AI ...167

Rozdział 8. Iluzja zaufania: „zatrute" dane, zmodyfikowane modele i ryzyko w łańcuchu dostaw powiązanym z technologią AI 197

Rozdział 9. Budowanie bezpiecznych łańcuchów dostaw modeli z udziałem zespołów ds. bezpieczeństwa wspieranych przez technologię AI 229

CZĘŚĆ III. ZACHOWANIE RÓWNOWAGI: ZAUFANIE, RYZYKO I ODPOWIEDZIALNOŚĆ

Rozdział 10. Ludzkie wartości i decyzje maszyn: etyka technologii AI w bezpieczeństwie 259

Rozdział 11. Przewaga człowieka: przywództwo, uczenie się i przyszłość bezpieczeństwa opartego na technologii AI 281

Kup książkę Poleć książkę

Rozdział 1

Era adaptacyjnych adwersarzy: wpływ sztucznej inteligencji na równowagę sił w bezpieczeństwie

Nowe tempo konfliktu: sztuczna inteligencja po obu stronach cybernetycznego pola bitwy

Każda fala technologii zbliżała nas do siebie: do siebie nawzajem i do naszych dzieł. Telefon przenosi nasze głosy na odległość. Internet przenosi nasze słowa w czasie. Technologia chmury przenosi nasze dane między systemami. Sztuczna inteligencja przenosi obecnie nasz osąd ponad granicę między człowiekiem a maszyną.

Każda zmiana przesuwała granice ryzyka i zaufania. Każda zmuszała obrońców do ponownego przemyślenia, co znaczy zabezpieczyć coś, co nieustannie zmienia kształt. Jednakże żadna z nich nie zmieniła tempa konfliktu tak bardzo jak sztuczna inteligencja (AI — *Artificial Intelligence*). To, co kiedyś rozwijało się przez wiele dni lub miesięcy, teraz dzieje się w ciągu milisekund. Decyzje, które kiedyś wymagały analityków i operatorów, obecnie są podejmowane przez algorytmy działające z prędkością i w skali stanowiących wyzwanie dla tradycyjnego nadzoru.

Cyberbezpieczeństwo zawsze było strategiczną grą akcji i kontrakcji. Atakujący sondują, a obrońcy wprowadzają poprawki. Wykrywanie prowadzi do unikania. Zapobieganie prowadzi do adaptacji.

Kup książkę Poleć książkę

Różnica polega dziś na szybkości i skali. Nie obserwujemy już dwóch graczy wymieniających się ruchami na statycznej planszy. Sama plansza się powiększa, a automatyzacja w coraz większym stopniu kształtuje obie strony konfliktu.

Sztuczna inteligencja działa jako mnożnik siły dla każdego, kto ma z nią styczność.

Dla atakujących oznacza automatyzację i skalę: boty rozpoznawcze mapujące cele w kilka sekund, kod polimorficzny, który może zmieniać swoją strukturę w celu uniknięcia wykrycia, oraz kampanie inżynierii społecznej umożliwiające personalizowanie na dużą skalę.

Dla obrońców sztuczna inteligencja oznacza przewidywanie i orkiestrację. Filtruje ona szum, ustala priorytety tego, co ważne, i łączy kropki w oceanach danych. Pomaga zespołom ds. bezpieczeństwa przewidywać potencjalne ścieżki ataku, identyfikować wzorce, których nie dostrzegłby żaden człowiek, i reagować, zanim szkody się rozprzestrzenią. Odpowiednio ukierunkowana sztuczna inteligencja daje obrońcom to, czego zawsze potrzebowali najbardziej, czyli czas.

Jednakże każdej zalecie towarzyszy złożoność. Współczesne pole bitwy nie ogranicza się już do sieci i punktów końcowych. Obejmuje ono obecnie potoki danych, modele szkoleniowe i systemy decyzyjne, czyli niewidoczne warstwy logiki, które dokonują wyborów w naszym imieniu. Atakujący biorą na cel dane i logikę, które kształtują zachowanie modelu, a nie tylko podstawową infrastrukturę. Manipulują oni promptami, aby wpłynąć na wyniki modelu lub ominąć zabezpieczenia. Wykorzystują bardziej założenia wbudowane w algorytmy niż sam kod.

Jest to nowy stan równowagi w cyberbezpieczeństwie. Atakujący wprowadzają innowacje na dużą skalę, a dbający o obronę operacjonalizują zaufanie w taki sam sposób. Sztuczna inteligencja zmienia granice między atakiem a obroną, tworzeniem a kontrolą. Te same techniki, które wzmacniają model, mogą być również użyte do oszukania go. Ta sama automatyzacja, która napędza obronę, może wzmocnić atak.

I w ten sposób dochodzimy do głównej idei tej książki: sztuczna inteligencja jest zarówno mieczem, jak i tarczą. Zwiększa ludzki

Kup książkę Poleć książkę

potencjał po obu stronach pola bitwy. Prawdziwa odporność nie bierze się z wyboru jednej lub drugiej strony, lecz z opanowania obu.

W kolejnych rozdziałach przyjrzymy się tej dwoistości: temu, jak ją zrozumieć, jak projektować z myślą o niej i ostatecznie kierować za jej pośrednictwem. Ponieważ w tym nowym obszarze odmowa wzięcia udziału w grze prowadzi wyłącznie do przegranej.

Wspólny język: podstawowe terminy i pojęcia

Każda szybko rozwijająca się dziedzina wypracowuje swój własny żargon. Aby sensownie rozmawiać o tym, jak sztuczna inteligencja zmienia cyberbezpieczeństwo, niezbędny jest wspólny i sprawdzony słownik, czyli taki, z którego będą mogli korzystać zarówno liderzy, jak i praktycy. Nie są to definicje akademickie, lecz praktyczne, zaprojektowane tak, aby pomóc Ci zadawać trafniejsze pytania i szybciej rozpoznawać ryzyko.

Skrótowce takie jak AI są używane tak często, że ich znaczenie zaczyna się zacierać. Zacznijmy zatem od jasnego zdefiniowania tego pojęcia. Czym ono jest, a czym nie jest, i dlaczego ma znaczenie z punktu widzenia bezpieczeństwa.

Sztuczna inteligencja (AI)

Systemy AI są systemami, które wykonują zadania związane z ludzkim procesem poznawczym, czyli percepcją, rozpoznawaniem wzorców, podejmowaniem decyzji i uczeniem się. AI nie jest pojedynczą technologią, ale ekosystemem modeli, danych i pętli sprzężenia zwrotnego.

Oto dlaczego ma to znaczenie dla bezpieczeństwa i zaufania: skala sztucznej inteligencji szybko przekształca dane w decyzje. We właściwych rękach ta moc zapewnia ochronę, a w niewłaściwych jest groźna.

Uczenie maszynowe

Uczenie maszynowe (ML — *Machine Learning*) to zbiór metod, które poprawiają zachowanie modelu poprzez znajdowanie wzorców w danych, a nie poprzez sztywne określanie reguł.

Kup książkę Poleć książkę

Oto dlaczego to ma znaczenie: uczenie maszynowe sprawia, że środki obrony są adaptacyjne, ale wprowadza również nowe tryby awarii, gdy zmieniają się dane wejściowe, konteksty lub zachęty.

Duże modele językowe (LLM)

Duże modele językowe (LLM — *Large Language Model*) to sieci neuronowe trenowane na dużych partiach tekstu/kodu, które przewidują następny token w sekwencji, aby generować płynny język (i kod).

Oto dlaczego jest to ważne: modele LLM przyspieszają generowanie treści, przegląd kodu i klasyfikację. Mogą również umożliwiać skalowanie inżynierii społecznej lub generowanie warstwowego szkieletu exploitów.

Protokół MCP

Protokół MCP (*Model Context Protocol*) to otwarty standard łączenia aplikacji (i modeli) AI z zewnętrznymi narzędziami, źródłami danych i procesami za pośrednictwem spójnego i bezpiecznego interfejsu. Można go traktować jak standaryzowany sposób łączenia systemów opartych na sztucznej inteligencji z narzędziami i danymi, czyli tworzącą się warstwę interoperacyjności (coś w rodzaju „interfejsu USB-C dla aplikacji AI").

Oto dlaczego ma to znaczenie: korzystanie z narzędzi poszerza możliwości i powierzchnię ataku (autoryzacja, dostęp do danych, wykonywanie akcji). Protokół MCP centralizuje i standaryzuje te połączenia, które trzeba nadzorować i monitorować.

Potoki AI (system systemów)

Potoki AI reprezentują typową strukturę w rzeczywistych wdrożeniach: gromadzenie danych → wstępne przetwarzanie → wnioskowanie → przetwarzanie końcowe → narzędzia (wyszukiwanie, wykonywanie kodu) → moderacja danych wyjściowych.

Oto dlaczego ma to znaczenie: każdy etap jest zależnością i potencjalnym punktem awarii lub ataku („zatruwanie" danych, nadużywanie narzędzi, „wycieki"). Zabezpiecz potok, a nie tylko „model".

Kup książkę Poleć książkę

Sieci neuronowe

Sieci neuronowe to warstwowe modele *obliczeniowe* inspirowane biologicznymi neuronami, które przekształcają dane wejściowe w dane wyjściowe za pomocą wyuczonych parametrów.

Oto dlaczego ma to znaczenie: ich nieprzejrzystość komplikuje kwestię zapewnienia bezpieczeństwa. Bez instrumentacji i rejestrowania trudno jest wyjaśnić (lub zakwestionować) decyzje.

Dane treningowe

Dane treningowe zawierają przykłady, na których uczy się model. Kształtują one to, co model uogólnia, w tym jego uprzedzenia i martwe punkty.

Oto dlaczego ma to znaczenie: na uwadze trzeba mieć bezwartościowe dane na wejściu i na wyjściu, a ponadto „zatrute" na wejściu i z istniejącą „tylną furtką" na wyjściu. Integralność danych to integralność modelu.

Wagi modelu

Wagi modelu to wyuczone parametry reprezentujące to, co model przyswoił podczas trenowania.

Oto dlaczego ma to znaczenie: wagi stanowią zarówno własność intelektualną, jak i potencjalną powierzchnię ataku. Kradzież ujawnia możliwości. Manipulacja może zaszczepić trudne do wykrycia zachowania.

Prompty i ich wprowadzanie

Prompty to instrukcje lub kontekst dostarczane w celu ukierunkowania zachowania modelu. Wprowadzanie (lub wstrzykiwanie) promptów ma miejsce, gdy atakujący preparuje dane wejściowe (bezpośrednio lub pośrednio za pośrednictwem zewnętrznej zawartości), aby nadpisać instrukcje, wyprowadzić poufne dane lub zainicjować niezamierzone działania. Jest to obecnie ryzyko związane z generatywną sztuczną inteligencją (*GenAI — Generative Artificial Intelligence*)

Kup książkę Poleć książkę

monitorowane przez organizację OWASP z wiarygodnymi demonstracjami w świecie rzeczywistym.

Oto dlaczego ma to znaczenie: prompty funkcjonują jako rozwijająca się granica bezpieczeństwa. Traktuj niezaufaną zawartość z taką samą ostrożnością jak niezaufany kod.

Dryf modelu

Dryf modelu opisuje spadek wydajności, który następuje z upływem czasu z powodu zmiany rozkładów bazowych danych lub relacji między danymi wejściowymi i wyjściowymi (dryf koncepcji, dryf danych).

Oto dlaczego ma to znaczenie: model, który nie jest monitorowany, ponownie trenowany lub wycofywany, może po cichu zawieść w środowisku produkcyjnym.

Podstawowe zasady cyberbezpieczeństwa zinterpretowane na nowo

Sztuczna inteligencja może wydawać się czymś nowym, ale fundamenty dobrych zabezpieczeń nie uległy zmianie.

Te same ponadczasowe zasady, jakie chroniły systemy przez dziesięciolecia, nadal obowiązują. Musimy je tylko zinterpretować na nowo pod kątem świata, w którym stosowana technologia może się uczyć i dostosowywać, a czasem zaskakiwać nas.

Zasady są znajome: widoczność, nadzór, odpowiedzialność i zaufanie. Teren jest nowy: potoki danych, wagi modeli i wnioskowanie maszynowe. Misja pozostaje jednak ta sama, czyli zapewnienie bezpieczeństwa systemów i nienaruszalności ludzkich wartości.

Triada CIA wymyślona na nowo

Każdy specjalista ds. bezpieczeństwa zna następujące trzy filary: poufność, integralność i dostępność (CIA — *Confidentiality, Integrity, Availability*). Stanowią one fundament cyberbezpieczeństwa, zanim jeszcze słowo „cyber” w ogóle powstało (czy ktoś pamięta termin InfoSec?), i nadal — w erze sztucznej inteligencji — są naszymi drogowskazami.

Kup książkę Poleć książkę

- Poufność oznacza obecnie ochronę danych treningowych, wag modeli i interfejsów API wnioskowania. Kto ma dostęp do zbiorów danych, które kształtują sposób rozumienia przez model? Co zapobiega „wyciekowi" informacji poufnych w ramach danych wyjściowych? Szyfrowanie, anonimizacja i solidna kontrola dostępu pozostają tak samo ważne jak zawsze.
- Integralność gwarantuje, że to, czego uczy się sztuczna inteligencja i jak się ona zachowuje, nie zostało zmanipulowane. Oznacza to haszowanie i podpisywanie modeli, śledzenie pochodzenia danych oraz ochronę przed „zatruwaniem" lub manipulacją adwersarialną.
- Dostępność zapewnia niezawodność w przypadku obciążenia. Systemy z modelem AI zawodzą na nowe sposoby, takie jak przeciążenie modelu, zmniejszona dokładność lub przestoje wynikające z zależności, ale rozwiązaniem jest klasyczna odporność, czyli redundancja, ograniczanie przepustowości i bezpieczne przełączanie awaryjne.

Triada się nie zmieniła. Po prostu „klejnoty koronne" zostały przeniesione z baz danych do modeli.

Obrona w głąb

Od lat twierdzono, że w dziedzinie bezpieczeństwa nie ma cudownego rozwiązania. Jest to bardziej prawdziwe niż kiedykolwiek w przypadku systemów sztucznej inteligencji, w których powierzchnia ataku obejmuje obecnie szkolenie, wdrażanie i interakcję z użytkownikiem.

Wielowarstwowe metody obrony oznaczają obecnie zabezpieczanie potoków danych, izolowanie środowisk dla modeli, kontrolowanie dostępu do narzędzi służących do dostrajania, monitorowanie wyników pod kątem anomalii oraz dodawanie mechanizmu wyłączników, aby zapobiec niekontrolowanej automatyzacji. Każda warstwa przekształca potencjalny exploit w niedogodność, a nie w katastrofę.

Obrona w głąb przypomina nam, że złożoność wymaga redundancji, a nie ślepego zaufania.

Kup książkę Poleć książkę

Modelowanie zagrożeń w erze maszyn

W ramach tradycyjnego modelowania zagrożeń stawiane są następujące pytania: Co chronimy? Kto może to zaatakować? W jaki sposób może mu się to udać? W przypadku sztucznej inteligencji pytania te ewoluują. Zasoby są nowe: dane, modele, prompty i potoki. Oprócz tradycyjnych podmiotów stanowiących zagrożenie, które dostosowują swoje techniki, wśród przeciwników pojawiają się nowe role, takie jak osoby „zatruwające" modele, wprowadzające prompty, pobierające dane i w szkodliwy sposób dostrajające modele. Dobre praktyki nadal zaczynają się tak samo: narysowanie diagramu, nazwanie ryzyk i przetestowanie swoich założeń. Obecnie jednak w dyskusji uwzględniamy odwracanie modelu, wprowadzanie promptów i „wyciek" danych.

Gdy już rozumiemy, w jaki sposób atakujący mogą manipulować tym, co dostrzega sztuczna inteligencja, możemy bronić tego, co będzie efektem podjętej przez nią decyzji.

Reagowanie na incydenty i podejście typu „czarna skrzynka"

Gdy w systemie z modelem AI dzieje się coś złego (nieoczekiwana odpowiedź, nagła stronniczość, dziwne zachowanie), trzeba prześledzić zapis informacji aż do ich początku. Właśnie z tego powodu rejestrowanie i monitorowanie jest obecnie tak samo niezbędne jak w początkach istnienia dzienników serwerów i rejestrów dostępu.

Systemy z modelem AI zyskują na solidnym rejestrowaniu i obserwowalności, czyli podejściu w stylu lotniczej „czarnej skrzynki".

Powinniśmy wiedzieć: Jakich danych użył model? Kto je zmienił i kiedy? Jakie prompty spowodowały określone wyniki? Ścieżki audytu i obserwowalność to nie tylko kwestia zgodności z przepisami. Jest to sposób, w jaki rozumiemy incydenty, które rozwijają się z szybkością maszyn. Nowoczesne plany reagowania na incydenty muszą obejmować procedury dotyczące dryfu modelu, „wycieków" opartych na promptach i „zatruwania" danych. Gdy modele działają nieprawidłowo, wycofywanie zmian i ponowne trenowanie stanowią nowe zarządzanie poprawkami.

Kup książkę Poleć książkę

Dostęp do danych i nadzór nad nimi

Nadzór sprowadza się do tego, kto jest w posiadaniu kluczy i kto monitoruje te osoby. Każda porcja danych użyta do trenowania modelu AI niesie ze sobą ryzyko, niezależnie od tego, czy są to dane wrażliwe, niedokładne, czy nieaktualne. Dobry nadzór oznacza wyraźne etykietowanie danych, ograniczanie sposobu ich wykorzystania oraz śledzenie tego, dokąd trafiają po opuszczeniu Twojej firmy. Oznacza to również zarządzanie ekspozycją na ryzyko po stronie dostawców i partnerów.

W przypadku sztucznej inteligencji Twój profil ryzyka nie ogranicza się do tego, co zbudujesz. Obejmuje on każdy model i usługę, z którą się łączysz. Zaufanie nie może być tylko przeczuciem. Musi ono zostać zweryfikowane.

„Człowiek w pętli"

Nawet najbardziej zaawansowane systemy z modelem AI wciąż wymagają oceny dokonanej przez człowieka. Maszyny są wyjątkowe pod względem wykrywania wzorców, a ludzie wyróżniają się umiejętnością rozumienia kontekstu, empatią i konsekwencją. „Człowiek w pętli" (HITL — *Human-in-the-Loop*) nie stanowi wąskiego gardła. Jest to zabezpieczenie.

Możemy zlecić modelowi AI badania, analitykę i generowanie opcji, lecz nigdy odpowiedzialność.

„Człowiek w pętli" to element, w przypadku którego ktoś zadaje następujące pytania: „Czy to ma sens?", „Czy to jest sprawiedliwe?", „Czy to jest bezpieczne?". Nadzór ze strony człowieka sprawia, że automatyzacja jest zgodna z ludzkimi wartościami. Jest to różnica między szybkością a lekkomyślnością, między autonomią a odpowiedzialnością.

Ewidencja i własność

Nie możesz chronić czegoś, jeśli nie wiesz, że to posiadasz. Każda organizacja wymaga aktualnej i dokładnej ewidencji swoich systemów z modelem AI. Co one robią? Jakich danych używają? Kto je zbudował

Kup książkę Poleć książkę

i kto odpowiada za ich wyniki? Taka ewidencja stanowi podstawę skutecznego wyjaśniania, nadzoru oraz zarządzania ryzykiem.

Bez niej nadzór i zarządzanie ryzykiem są niemożliwe.

Stara mądrość, nowy teren

Sztuczna inteligencja może zmieniać narzędzia, ale nie zmienia fundamentów. Rejestrowanie, nadzór, kontrola sprawowana przez człowieka i ewidencja pozostają fundamentami skutecznego cyberbezpieczeństwa. Po prostu rozszerzamy je na systemy, które się uczą, wnioskują i działają.

Najlepszą obroną nie jest zaczynanie od nowa, lecz budowanie na tym, co już jest. Oznacza to stosowanie przejrzystości, odpowiedzialności i weryfikacji w nowych kontekstach. Gdy organizacje integrują sztuczną inteligencję, podstawowe zasady nadal obowiązują. Ustal, czym dysponujesz, monitoruj jej zachowanie, bierz odpowiedzialność za wyniki i zapewnij, by kontrolę sprawował człowiek.

Mój apel o możliwość wyjaśniania i transparentność

Sztuczna inteligencja na nowo definiuje sposób podejmowania decyzji. Oznacza to, że definiuje też na nowo nasze wyobrażenie o tym, jak wygląda zaufanie. W przypadku cyberbezpieczeństwa zawsze uważałam, że zaufanie zależy od przejrzystości, czyli wiedzy o tym, kto co zrobił, kiedy i dlaczego.

Sztuczna inteligencja komplikuje jednak to równanie. Gdy system generuje kod, zatwierdza dostęp lub zaleca środki zaradcze, potrzebujemy czegoś więcej niż tylko dokładności. Niezbędna jest odpowiedzialność. Z tego powodu apeluję, aby możliwość wyjaśniania i przejrzystość stały się kolejnym standardem dojrzałości cybernetycznej.

Dlaczego jest to takie ważne?

Zaufanie wymaga wglądu w to, jak sztuczna inteligencja podejmuje decyzje. Bez tego nie możemy audytować, nadzorować ani ulepszać systemów, które w coraz większym stopniu nas kontrolują. Nieprzejrzyste

Kup książkę Poleć książkę

algorytmy nie tylko podważają zaufanie do technologii. Osłabiają również zaufanie do organizacji, które ją wdrażają.

Możliwość wyjaśniania to sposób, w jaki zamieniamy automatyzację w pewność. Dotyczy to tego, jak przekształcamy „model tak stwierdził" w dowody, które możemy ocenić i na których podstawie możemy działać. Musimy oceniać przejrzystość na kilka następujących sposobów:

- **Przejrzystość techniczna.** Jak to działa? Nie trzeba dostrzec każdej wagi ani neuronu. Musimy zrozumieć wystarczająco dużo na temat projektu modelu, danych i ścieżek decyzyjnych, aby wykryć stronniczość, błąd lub manipulację.
- **Przejrzystość operacyjna.** Skąd to się wzięło? Musimy być w stanie prześledzić rodowód modelu. Jakie dane zostały użyte do trenowania go? Kto go dostrajał? Kiedy był ostatnio aktualizowany i jaka wersja działa w środowisku produkcyjnym? Pochodzenie, kontrola wersji i zarządzanie zmianami nie są już opcjonalne. Są niezbędne do bezpiecznego wdrażania.
- **Przejrzystość etyczna.** Kto ponosi odpowiedzialność? Każda zautomatyzowana decyzja nadal należy do człowieka lub organizacji. Możemy delegować analizę maszynom, ale nie możemy delegować odpowiedzialności. Ramy nadzoru muszą określać odpowiedzialność za szkody, stronniczość i niewłaściwe użycie. Przejrzystość etyczna to sposób, w jaki dostosowujemy technologię do ludzkich wartości, zamiast zlecać ocenę kodowi.

Mechanizmy w praktyce

Potrzeba możliwości wyjaśniania nie jest tylko teoretyczna. Odpowiednie narzędzia już istnieją. Trzeba jedynie używać ich z zachowaniem dyscypliny i zamysłem.

- Karty modeli dokumentują przeznaczenie modelu, źródła danych, charakterystyki wydajności i znane ograniczenia.
- Dzienniki audytu rejestrują kluczowe dane wejściowe i wyjściowe oraz aktualizacje modelu w celu zapewnienia odpowiedzialności i możliwości sprawdzenia.

Kup książkę Poleć książkę

- Śledzenie pochodzenia pozwala na śledzenie zbiorów danych, wersji modeli i artefaktów w trakcie całego cyklu życia procesu projektowania i wdrażania.
- Testowanie odtwarzalności pozwala weryfikować, czy w tych samych warunkach zachowanie modelu pozostaje spójne i jest możliwe do wyjaśnienia.

Razem mechanizmy te stanowią fundament obserwowalności sztucznej inteligencji, czyli zdolności do wyjaśniania, śledzenia i weryfikowania tego, co i dlaczego system zrobił.

Punkty zaczepienia ram

Organy regulacyjne i opracowujące standardy podążają w tym samym kierunku. Ustawa Unii Europejskiej dotycząca sztucznej inteligencji (*AI Act*) wymaga dokumentacji opartej na ryzyku, nadzoru ze strony człowieka i środków transparentności w kontekście systemów z modelem AI wysokiego ryzyka. Ramy zarządzania ryzykiem w przypadku sztucznej inteligencji (*AI Risk Management Framework*) opracowane przez agencję NIST definiują transparentność i możliwości wyjaśniania jako podstawowe cechy sztucznej inteligencji godnej zaufania, kładąc nacisk na kwestię dokumentacji i odpowiedzialności. Norma ISO/IEC 42001 ustanawia ramy systemu zarządzania na potrzeby nadzoru sztucznej inteligencji, wprowadzając mechanizmy kontroli w zakresie odpowiedzialności, możliwości przeprowadzania audytu i ciągłego doskonalenia. Ramy te różnią się zakresem, ale łączy je wspólna prawda, z którą się zgadzam: nie można zabezpieczyć tego, czego nie można wyjaśnić.

Sztuczna inteligencja umożliwiająca wyjaśnianie to nie jest termin marketingowy, lecz kolejny etap dojrzałości cyberbezpieczeństwa.

Podobnie jak rejestrowanie, wskaźniki i reagowanie na incydenty stały się standardową praktyką w przypadku sieci i aplikacji, tak możliwość wyjaśniania stanie się obowiązkowa w odniesieniu do sztucznej inteligencji. Organizacje, które wiodą prym w zakresie transparentności, zyskają zaufanie użytkowników i przewagę operacyjną. Przyszłość należy do systemów, które potrafią udowodnić swoją integralność. Zapewnienie możliwości wyjaśniania to sposób, w jaki

Kup książkę Poleć książkę

utrzymujemy kontrolę człowieka nad maszynami, a ponadto w jaki przekształcamy nieprzejrzyste systemy w infrastrukturę oferującą rozliczanie i zapewniamy, że postępowi w dziedzinie inteligencji towarzyszy postęp w kwestii zaufania.

Trwała przewaga ludzi: ciekawość, odwaga i zmiana

Każda era postępu zaczyna się od nowego narzędzia i wyobraźni, by z niego skorzystać. Kiedy pierwsi ludzie wyciosali z krzemienia ostrze, zwiększyli zasięg swoich działań. Kompas dał odkrywcom odwagę do przemierzania oceanów. Prasa drukarska zwiększyła dostęp do wiedzy. Tranzystor skompresował pomieszczenia z maszynami obliczeniowymi do rozmiaru dłoni.

Teraz do tej tradycji dołącza sztuczna inteligencja. Jest to nasz najnowszy instrument. Nie jest to konkurencyjny umysł, ale czynnik zwiększający nasze możliwości. Technologia AI rozszerza ludzką percepcję i możliwość rozpoznawania wzorców. Przyspiesza naszą zdolność łączenia kropek i dostrzegania tego, co dotąd ukrywało się w szumie. W dziedzinie cyberbezpieczeństwa oznacza to, że obrońcy mogą szybciej wykrywać ataki, analizować więcej danych i reagować w czasie wynoszącym kilka sekund, a nie godzin.

Jednakże, tak jak każde narzędzie wcześniej, technologia AI odzwierciedla intencje. Ogień może oświetlać lub niszczyć. Kod może wzmacniać lub wykorzystywać podatności. Inteligencja, czy to naturalna czy sztuczna, może zostać nakierowana ku odporności lub ruinie. Celem nie jest przechytrzenie maszyn, ale nauczenie się, jak z nimi współpracować. Służy to zbudowaniu pętli sprzężenia zwrotnego, w ramach których ludzki osąd kształtuje precyzję maszyny.

Właściwie używana technologia AI nie zastępuje ludzkiego myślenia, lecz je wzmacnia. Zamienia ona dane w przewidywanie, a przewidywanie w odporność. Każdy krok naprzód wymagał od nas adaptacji, czyli przedefiniowania siebie względem używanych narzędzi. Wyzwaniem nigdy nie była sama technologia, ale nasza chęć do nauki.

Uczenie się to najstarsza umiejętność służąca przetrwaniu ludzkości. To ono zamienia niepewność w innowację, a ryzyko w szansę.

Kup książkę Poleć książkę

Aby prosperować w erze technologii AI, potrzebujemy tego samego ducha ciekawości — nastawienia początkującego, który z zadowoleniem przyjmuje zmiany i widzi możliwości tam, gdzie inni dostrzegają komplikacje. Systemy, które budujemy, będą odzwierciedlać nastawienie ich twórców: będą otwarte lub zamknięte, adaptacyjne lub sztywne, przejrzyste lub nieprzejrzyste.

I to prowadzi mnie do Ciebie.

Czytasz to, ponieważ zdecydowałeś się uczyć i lepiej zrozumieć dziedzinę, która zmienia się szybciej niż sama pewność. Wybór ten ma znaczenie. Łatwo poczuć się przytłoczonym zmianą. Zaangażowanie się w nią wymaga odwagi. Fakt, że to robisz — czytasz, odkrywasz i rozwijasz się — jest pierwszym i najważniejszym aktem odporności. Ponieważ w świecie, w którym technologia ewoluuje szybciej niż zasady, ciekawość jest najpotężniejszym środkiem kontroli bezpieczeństwa, jakim dysponujemy.

Narzędzia będą stawać się coraz mądrzejsze. Ataki będą coraz szybsze. Jednakże przewaga — ta ciągła przewaga człowieka — zawsze będzie należeć do tych, którzy wciąż się uczą.

Kup książkę Poleć książkę


PROGRAM
PARTNERSKI
GRUPY HELION
1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ
Zmień swoją stronę WWW w działający bankomat!
Dowiedz się więcej i dołącz już dzisiaj!
http://program-partnerski.helion.pl
GRUPA
Helion

AI TO MIECZ I TARCZA — NAUCZ SIĘ WŁADAĆ OBOMA!

Sztuczna inteligencja i cyberbezpieczeństwo idą w parze — i to coraz ściślej. W dobie rosnących zagrożeń cyfrowych, autonomicznych ataków i błyskawicznie ewoluujących technik hakerskich AI stała się zarówno kluczowym elementem arsenału atakujących, jak i tarczą specjalistów od obrony. Temat ten niezmiennie dominuje w dyskusjach branżowych, raportach bezpieczeństwa i strategiach IT największych organizacji na świecie. Książka AI w cyberbezpieczeństwie wpisuje się w ten nurt, oferując kompleksowe i praktyczne spojrzenie na przecięcie tych dwóch dziedzin — bez zbędnego szumu informacyjnego, za to z solidnym fundamentem merytorycznym.

Autorka prowadzi czytelnika przez cały ekosystem zagrożeń i mechanizmów obronnych opartych na AI. Analizuje, jak sztuczna inteligencja przyspiesza rozpoznanie, wzmacnia złośliwe oprogramowanie i industrializuje phishing, a jednocześnie pokazuje, jak specjaliści od obrony mogą zastosować uczenie maszynowe do wykrywania anomalii, automatyzacji reagowania na incydenty i zabezpieczania łańcuchów dostaw modeli. Książka nie pomija trudnych pytań o etykę, stronniczość algorytmów, odpowiedzialność za decyzje maszyn i nadzór człowieka — traktuje je jako operacyjną konieczność, nie filozoficzny luksus.

Najważniejsze zagadnienia:

- Jak AI automatyzuje rozpoznanie i personalizuje ataki socjotechniczne na niespotykaną skalę
- Polimorficzne i metamorficzne złośliwe oprogramowanie — kiedy kod zaczyna się uczyć i adaptować
- Deepfake, klonowanie głosu i hiperukierunkowany phishing jako nowe narzędzia cyberprzestępców
- Wykrywanie anomalii behawioralnych i reagowanie na incydenty wspomagane przez AI
- Bezpieczeństwo łańcucha dostaw modeli AI — „zatruwanie" danych, manipulacja wagami, wstrzykiwanie promptów
- Etyka AI w bezpieczeństwie: stronniczość, przejrzystość, nadzór człowieka i odpowiedzialność za decyzje maszyn

CAROLINE WONG to uznana strateżka do spraw cyberbezpieczeństwa z ponad 20-letnim doświadczeniem w obszarach GRC, bezpieczeństwa produktów i zarządzania na poziomie wykonawczym. Pełni funkcję dyrektorki w Teradata, jako wykładowczyni i konsultantka w IANS Research doradza organizacjom z listy Fortune 500. Jest autorką nagrodzonej książki *Security Metrics. A Beginner's Guide* (Cybersecurity Canon Hall of Fame, 2022). Prowadzi trzy podcasty branżowe: *Humans of InfoSec*, *The AI Security Edge* i *Wake Up*. Wcześniej pracowała między innymi w eBay, Zynga i Symantec.

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-4231-8
9 788328 942318
Cena: 89,00 zł
WILEY