

Spis treści |

O autorze	17
O recenzentach	17
Zastrzeżenie	18
Wstęp	19

CZĘŚĆ 1. Wprowadzenie do adversarialnej sztucznej inteligencji

ROZDZIAŁ 1

Wprowadzenie do sztucznej inteligencji	27
Podstawy sztucznej inteligencji i uczenia maszynowego	28
Rodzaje uczenia maszynowego i jego cykl życia	29
Kluczowe algorytmy uczenia maszynowego	32
Sieci neuronowe i uczenie głębokie	33
Narzędzia deweloperskie do programowania uczenia maszynowego	35
Podsumowanie	36
Dodatkowe publikacje	37

ROZDZIAŁ 2

Przygotowanie adversarialnego laboratorium	38
Wymagania techniczne	39
Konfiguracja środowiska programistycznego	39
Instalacja Pythona	39
Tworzenie wirtualnego środowiska	40
Instalowanie pakietów	41
Powiązanie środowiska wirtualnego z notatnikami Jupytera	41
Weryfikowanie instalacji	41
Praktyczne podstawy uczenia maszynowego	42
Proste sieci neuronowe	45
Tworzenie docelowej usługi AI z użyciem sieci CNN	46
Konfiguracja i gromadzenie danych	46

Eksploracja danych	47
Wstępne przetwarzanie danych	47
Wybór algorytmu i tworzenie modelu	48
Trenowanie modelu	48
Ewaluacja modelu	49
Wdrażanie modelu	50
Usługa wnioskowania	50
Tworzenie rozwiązań ML na dużą skalę	51
Google Colab	52
AWS SageMaker	52
Usługi Azure Machine Learning	52
Lambda Labs Cloud	53
Podsumowanie	53

ROZDZIAŁ 3

Bezpieczeństwo a adversarialna sztuczna inteligencja	55
Wymagania techniczne	56
Podstawy bezpieczeństwa	56
Modelowanie zagrożeń	57
Ryzyka i środki zaradcze	57
DevSecOps	58
Zabezpieczenie naszego adversarialnego laboratorium	58
Bezpieczeństwo hosta	60
Ochrona sieci	63
Uwierzytelnianie	65
Ochrona danych	66
Kontrola dostępu	69
Zabezpieczanie kodu i artefaktów	70
Bezpieczeństwo kodu	71
Zabezpieczanie zależności przez wyszukiwanie podatności	71
Skanowanie sekretów	74
Zabezpieczanie notatników Jupytera	74
Zabezpieczanie modeli przed złośliwym kodem	76
Integracja z potokami DevSecOps i MLOps	77
Omijanie zabezpieczeń za pomocą adversarialnej sztucznej inteligencji	77
Nasz pierwszy adversarialny atak	78
Tradycyjne cyberbezpieczeństwo a adversarialna sztuczna inteligencja	80
Adversarialna sztuczna inteligencja	81
Podsumowanie	81

CZĘŚĆ 2. Atakowanie modeli ML

ROZDZIAŁ 4

Ataki zatruwające	85
Podstawy ataków zatruwających	86
Definicja i przykłady	86
Rodzaje ataków zatruwających	87
Przykłady ataków zatruwających	88
Dlaczego to takie ważne?	89
Przygotowanie prostego ataku zatruwającego	90
Tworzenie zatrutych próbek	90
Ataki zatruwające typu backdoor	94
Tworzenie wyzwalaczy backdoorowych za pomocą narzędzi ART.	98
Zatrutowanie danych z użyciem frameworku ART.	102
Ataki backdoorowe z ukrytym wyzwalaczem	103
Ataki typu clean-label	104
Zaawansowane ataki zatruwające	106
Zapobieganie i obrona	107
Obrona cyfrowych twierdz przy użyciu MLOps	107
Wykrywanie anomalii	108
Testy odporności na zatrutowanie	109
Zaawansowana ochrona przed zatruciem realizowana z użyciem narzędzi ART.	110
Trening adversarialny	112
Budowanie strategii obronnej	112
Podsumowanie	113

ROZDZIAŁ 5

Manipulowanie modelami za pomocą koni trojańskich i przeprogramowywania	114
Wstrzykiwanie backdoorów za pomocą serializacji pickle'owej	115
Przebieg ataku	115
Obrona i środki zaradcze	117
Wstrzykiwanie koni trojańskich za pomocą kerasowych warstw lambda ...	118
Przebieg ataku	118
Obrona i środki zaradcze	122
Warstwy niestandardowe z końmi trojańskimi	124
Przebieg ataku	124
Obrona i środki zaradcze	126

Wstrzykiwanie ładunku neuronowego	127
Przebieg ataku	128
Obrona i środki zaradcze	131
Atakowanie brzegowej sztucznej inteligencji	132
Aplikacja mobilna ImRecS na Androida	132
Przebieg ataku	133
Obrona i środki zaradcze	134
Przejmowanie kontroli nad modelem	137
Wprowadzenie kodu konia trojańskiego	137
Przeprogramowanie modelu	138
Podsumowanie	139

ROZDZIAŁ 6

Ataki na łańcuch dostaw a adversarialna sztuczna inteligencja 140

Tradycyjne zagrożenia dla łańcucha dostaw a sztuczna inteligencja	141
Zagrożenia związane z przestarzałymi i podatnymi na atak komponentami	141
Ryzyka wynikające z zależności AI od danych pobieranych na żywo	143
Zabezpieczanie AI przed podatnymi na atak komponentami	145
Zaawansowane zabezpieczenia — akceptowanie jedynie zatwierdzonych pakietów	150
Konfiguracja klienta do obsługi prywatnych repozytoriów PyPI	151
Dodatkowe zabezpieczenia prywatnego repozytorium PyPI	152
Korzystanie ze specyfikacji SBOM	153
Ryzyka związane z łańcuchem dostaw w kontekście AI	155
Uczenie transferowe jako broń obosieczna	155
Zatrutowanie modelu	156
Manipulowanie modelami	161
Sprawdzanie pochodzenia wstępnie wytrenowanych modeli i nadzorowanie ich	163
MLOps a prywatne repozytoria modeli	164
Zatrutowanie danych	168
Ryzyka związane z łańcuchem dostaw	168
Zatrutowanie danych, aby wpłynąć na wyniki analizy sentymentu	169
Obrona i środki zaradcze	171
Specyfikacje komponentów oprogramowania (SBOM) dla AI/ML	172
Podsumowanie	172

CZĘŚĆ 3. Ataki na implementacje AI

ROZDZIAŁ 7

Ataki unikowe na implementacje AI	177
Podstawy ataków unikowych	177
Znaczenie znajomości ataków unikowych	178
Techniki rozpoznawcze w atakach unikowych	179
Perturbacje i techniki ataków unikowych w kontekście obrazów	182
Scenariusze ataków unikowych	183
Jednoetapowa perturbacja z użyciem metody FGSM	184
Podstawowa metoda iteracyjna (BIM)	187
Atak z mapą istotności wyznaczonej metodą Jacobiego (JSMA)	188
Atak Carliniego i Wagnera (C&W)	190
Rzutowanie gradientu (PGD)	192
Łatki adversarialne — łączenie cyfrowych i fizycznych technik unikania	193
Ataki unikowe w dziedzinie NLP — atak na model BERT z użyciem biblioteki TextAttack	195
Scenariusz ataku — analiza sentymentu	195
Przykład ataku	195
Scenariusz ataku — wnioskowanie w języku naturalnym	197
Przykład ataku	197
Uniwersalne perturbacje adversarialne (UAP)	198
Scenariusz ataku	199
Przykład ataku	199
Ataki czarnoskrzynkowe oraz ich transferowalność	200
Scenariusz ataku	201
Przykład ataku	201
Obrona przed atakami unikowymi	202
Przegląd strategii obronnych	202
Trening adversarialny	203
Wstępne przetwarzanie danych wejściowych	204
Techniki wzmacniania modelu	205
Zespoły modeli	207
Certyfikowane mechanizmy obronne	208
Podsumowanie	209

ROZDZIAŁ 8

Ataki na prywatność — kradzież modeli	210
Charakterystyka ataków na prywatność	210
Wykradanie modeli podczas ataków ekstrakcyjnych	211
Ekstrakcja równoważna funkcjonalnie	212
Ataki ekstrakcyjne oparte na uczeniu	214
Generatywne ataki ekstrakcyjne typu uczeń-nauczyciel (ataki destylacyjne)	219
Przykładowy atak na model CIFAR-10 CNN	222
Obrona i środki zaradcze	227
Środki prewencyjne	227
Mechanizmy wykrywające	231
Ustalanie stanu własności modelu oraz jej odzyskiwanie	232
Podsumowanie	235

ROZDZIAŁ 9

Ataki na prywatność — kradzież danych	236
Istota ataków polegających na inwersji modelu	236
Typy ataków inwersyjnych	238
Wykorzystanie poziomów pewności modelu	238
Inwersja modelu wspomagana sieciami GAN	240
Przykład ataku inwersyjnego	246
Istota ataków wnioskowania	248
Ataki wnioskowania o atrybutach	248
Metaklasyfikatory	249
Wnioskowanie wspomagane zatruciem	249
Przykład ataku wnioskowania o atrybutach	251
Ataki wnioskowania o przynależności	253
Statystyczne wartości progowe wycieku danych z modeli ML	254
Atak transferu wiedzy z wykorzystaniem samych etykiet	255
Ślepe ataki wnioskowania o przynależności	255
Ataki białoskrzynkowe	256
Przykładowy atak wnioskowania o przynależności z użyciem narzędzi ART.	259
Podsumowanie	260

ROZDZIAŁ 10

Zachowanie prywatności w rozwiązaniach AI	262
Zachowanie prywatności w rozwiązaniach ML i AI	263
Prosta anonimizacja danych	264

Zaawansowana anonimizacja	268
K-anonimowość	268
Anonimizacja a dane geolokalizacyjne	270
Anonimizacja formatów multimedialnych	272
Prywatność różnicowa	278
Uczenie federacyjne	281
Uczenie dzielone	282
Zaawansowane opcje szyfrowania wspomagające ochronę prywatności	
w uczeniu maszynowym	283
Bezpieczne obliczenia wielostronne	283
Szyfrowanie homomorficzne	285
Praktyczne zastosowanie zaawansowanych technik szyfrowania	
w rozwiązaniach ML	287
Stosowanie technik zapewniania prywatności w uczeniu maszynowym	289
Podsumowanie	290

CZĘŚĆ 4. Generatywna sztuczna inteligencja a ataki adwersarialne

ROZDZIAŁ 11

Generatywna sztuczna inteligencja — nowy front walki	293
Krótkie wprowadzenie do generatywnej sztucznej inteligencji	294
Krótka historia rozwoju generatywnej sztucznej inteligencji	294
Technologie generatywnej sztucznej inteligencji	295
Stosowanie generatywnych sieci adwersarialnych	298
Tworzenie sieci GAN od podstaw	299
Sieci WGAN i niestandardowe funkcje strat	307
Korzystanie ze wstępnie wytrenowanych modeli GAN	308
Pix2Pix	308
CycleGAN	309
Pix2PixHD	309
PGGAN	310
BigGAN	310
StarGAN v2	311
Seria StyleGAN	311
Podsumowanie	312

ROZDZIAŁ 12**Wykorzystywanie sieci GAN do deepfake'ów****i ataków adversarialnych 313**

Wykorzystanie GAN-ów do tworzenia deepfake'ów oraz ich wykrywania 313

Generowanie przekonujących fałszywych zdjęć za pomocą modeli StyleGAN 314

Wykorzystanie sieci GAN do tworzenia prostych deepfake'ów na podstawie istniejących zdjęć 318

Wprowadzanie ukierunkowanych zmian w istniejącym obrazie 320

Syntezowanie obrazów za pomocą modelu Pix2PixHD 321

Sfałszowane filmy i animacje 324

Inne techniki tworzenia deepfake'ów 325

Deepfaki dźwiękowe 329

Wykrywanie deepfake'ów 330

Zastosowanie GAN-ów w cyberatakach i bezpieczeństwie ofensywnym 333

Omijanie systemów weryfikacji twarzy 333

Oszukiwanie biometrycznych mechanizmów uwierzytelniających 335

Łamanie haseł z użyciem GAN-ów 337

Omijanie mechanizmów detekcji złośliwego oprogramowania 340

GAN-y w kryptografii i steganografii 342

Generowanie ładunków dla ataków sieciowych z użyciem GAN-ów 344

Generowanie ładunków dla ataków adversarialnych 345

Mechanizmy obronne i środki zaradcze 348

Zabezpieczanie sieci GAN 348

Ataki adversarialne wspomagane sieciami GAN 349

Deepfaki, złośliwe treści i szerzenie dezinformacji 351

Podsumowanie 355

ROZDZIAŁ 13**Podstawy LLM w kontekście adversarialnej sztucznej inteligencji 377**

Krótkie wprowadzenie do dużych modeli językowych 357

Tworzenie aplikacji AI z użyciem dużych modeli językowych 359

„Hello LLM” w Pythonie 360

„Hello LLM” w LangChainie 365

Wprowadzanie własnych danych 366

Wpływ dużych modeli językowych na adversarialną sztuczną inteligencję 370

Podsumowanie 371

ROZDZIAŁ 14**Ataki adversarialne z użyciem promptów 372**

Adversarialne dane wejściowe i wstrzykiwanie promptów	373
Bezpośrednie wstrzykiwanie promptów	374
Zastępowanie promptów	375
Wstrzykiwanie stylu	378
Odgrywanie ról	378
Podszywanie się	382
Inne techniki jailbreakingowe	389
Zautomatyzowane wstrzykiwanie promptów z użyciem technik gradientowych	391
Ryzyko związane z wprowadzaniem własnych danych	393
Pośrednie wstrzykiwanie promptów	393
Wydobywanie danych za pomocą wstrzykiwania promptów	397
Eskalacja uprawnień za pomocą wstrzykiwania promptów	398
Zdalne wykonywanie kodu przez wstrzykiwanie promptów	399
Mechanizmy obronne i środki zaradcze	401
Mechanizmy obronne platformy LLM	402
Mechanizmy obronne na poziomie aplikacji	403
Podsumowanie	411

ROZDZIAŁ 15**Ataki zatruwające a modele LLM 412**

Zatrutowanie osadzeń w mechanizmie RAG	412
Scenariusze ataku	419
Zatrutowanie podczas generowania osadzeń	420
Bezpośrednie zatrutowanie osadzeń	427
Zaawansowane zatrutowanie osadzeń	428
Manipulowanie osadzeniami zapytań	432
Mechanizmy obronne i środki zaradcze	433
Ataki zatruwające proces dostrajania modeli LLM	435
Wprowadzenie do dostrajania modeli LLM	435
Scenariusze ataków zatruwających podczas dostrajania	438
Wektory ataku na proces dostrajania	440
Zatrucie bota ChatGPT-3.5 przez dostrajanie	441
Mechanizmy obronne i środki zaradcze chroniące przed zatrutowaniem procesu dostrajania	451
Podsumowanie	455

ROZDZIAŁ 16**Zaawansowane scenariusze z wykorzystaniem**

generatywnej sztucznej inteligencji	456
Ataki na łańcuch dostaw w kontekście LLM-ów	457
Publikowanie zatrutego LLM-u w serwisie Hugging Face	459
Publikowanie zmanipulowanego modelu LLM w serwisie Hugging Face	465
Inne zagrożenia związane z łańcuchem dostaw w kontekście modeli LLM	467
Mechanizmy obronne i środki zaradcze w kontekście łańcucha dostaw	468
Ataki na prywatność w kontekście modeli LLM	469
Ataki polegające na inwersji modelu i ekstrakcji danych treningowych w kontekście modeli LLM	470
Ataki wnioskowania na modele LLM	472
Klonowanie jednego modelu LLM przy użyciu drugiego	474
Mechanizmy obronne i środki zaradcze w kontekście ataków na prywatność	476
Podsumowanie	477

CZĘŚĆ 5. Zabezpieczanie AI przez projekt i praktyki MLSecOps

ROZDZIAŁ 17

Koncepcje Secure by Design i Trustworthy AI	481
Secure by Design — zabezpieczanie AI już na etapie projektowania	482
Budowanie biblioteki zagrożeń	485
Tradycyjne cyberzagrożenia	486
Ataki adversarialne	486
Ataki adversarialne specyficzne dla generatywnej AI	488
Ataki na łańcuch dostaw	489
Branżowe klasyfikacje zagrożeń dla AI	490
Porównania klasyfikacji zagrożeń dla AI	491
Porównanie z klasyfikacją NIST AI	491
Porównanie z klasyfikacją AI Exchange	495
Porównanie z klasyfikacją MITRE ATLAS	498
Modelowanie zagrożeń dla AI	501
Modelowanie zagrożeń w praktyce	502
Przykładowe rozwiązanie AI	502

Model zagrożeń dla systemu Enhanced FoodieAI 503

Ocena zagrożeń i ich priorytetyzacja 509

 Ocena ryzyka dla aplikacji Enhanced FoodieAI 514

Projektowanie i implementowanie zabezpieczeń 518

Testowanie i weryfikacja 534

Przesuwanie w lewo i osadzanie zabezpieczeń w cyklu życia AI 535

Eksploracja systemu 535

Więcej niż bezpieczeństwo — Trustworthy AI 538

Podsumowanie 540

ROZDZIAŁ 18

Zabezpieczanie AI przy użyciu strategii MLSecOps 541

 Konieczność wdrożenia praktyk MLSecOps 541

 Na drodze do frameworku MLSecOps 2.0 544

 Opcje orkiestracji MLSecOps 544

 Wzorce MLSecOps 547

 Budowanie podstawowej platformy MLSecOps 552

 MLSecOps w praktyce 557

 Pozyskiwanie i walidacja modelu 558

 Integracja MLSecOps z LLMOps 571

 Zaawansowane praktyki MLSecOps z użyciem SBOM-ów 574

 Podsumowanie 578

ROZDZIAŁ 19

Wzmacnianie bezpieczeństwa AI 579

 Wyzwania związane z bezpieczeństwem AI w przedsiębiorstwie 579

 Podstawy bezpieczeństwa sztucznej inteligencji w przedsiębiorstwie 582

 Ochrona sztucznej inteligencji przy użyciu zabezpieczeń korporacyjnych 584

 Bezpieczeństwo operacyjne AI 586

 Iteracyjne wzmacnianie bezpieczeństwa w przedsiębiorstwie 588

 Podsumowanie 588

Skorowidz 591