

**Hakowanie sztucznej inteligencji / redakcja naukowa Jerzy Surma ;
autorzy: Jerzy Surma, Piotr Filipkowski, Mariusz Rafało, Piotr Kuc,
Krzysztof Jagiełło, Kamil Frankowicz. – Wydanie I. – Warszawa, ©
copyright 2020**

Spis treści

Wstęp	9
Bibliografia	12
1. Wstęp do hakowania systemów uczących się	15
1.1. Wprowadzenie	16
1.2. Systemy uczące się	17
1.2.1. Definicja i rodzaje systemów uczących się	17
1.2.2. Zadanie klasyfikacji i uczenie nadzorowane	18
1.2.3. Ocena jakości klasyfikatora	19
1.2.4. Problemy budowania systemów uczących się	21
1.2.5. Potencjalne cele atakującego	22
1.3. Taksonomia ataków na systemy uczące się	23
1.3.1. Kryteria jakości ochrony informacji	23
1.3.2. Atak na integralność systemów nadzorowanych	25
1.3.2.1. Formalizacja ataku na integralność	25
1.3.2.2. Atak na proces budowania systemu	26
1.3.2.3. Atak na funkcjonujący system	27
1.3.3. Atak na integralność innych rodzajów systemów uczących się	29
Bibliografia	30
2. Przegląd reprezentatywnych ataków	33
2.1. Wprowadzenie	34
2.2. Zagrożenia dla systemów uwierzytelniania	35
2.3. Zagrożenia w systemach autonomicznych	40
2.4. Zagrożenia w systemach medycznych	45
2.5. Wnioski końcowe	49
Bibliografia	50
3. Wymiar biznesowy ataków na systemy uczące się	53
3.1. Wprowadzenie	54
3.2. Robotyzacja i automatyzacja procesów biznesowych	55
3.2.1. Robotyzacja procesów	55
3.2.2. Sztuczna inteligencja w robotyzacji procesów	57
3.3. Ryzyko operacyjne w procesach biznesowych	59
3.3.1. Problematyka ryzyka	60
3.3.2. Zarządzanie ryzykiem	60
3.3.3. Ryzyko w RPA działających z wykorzystaniem systemów uczących się	62

3.4. Zagrożenia związane z wykorzystaniem systemów uczących się w RPA	63
3.4.1. Wprowadzenie	63
3.4.2. Geneza ataków na systemy uczące się	65
3.4.3. Przykłady realnych zagrożeń	67
3.4.3.1. Uwagi wstępne	67
3.4.3.2. Przykład ataku infekcyjnego	68
3.4.3.3. Atak na automatyczny systemy w transakcji finansowych	69
3.4.3.4. Ataki na systemy rekomendacyjne	71
3.4.3.5. Inne zagrożenia	72
3.5. Zakończenie	74
Bibliografia	74
4. Studia przypadków	79
4.1. Atakowanie filtra antyspamowego wykorzystującego system uczący się	80
4.1.1. Charakterystyka problemu	80
4.1.1.1. Wprowadzenie	80
4.1.1.2. Definicja filtra antyspamowego	81
4.1.1.3. Problem filtrowania poczty elektronicznej w działalności biznesowej	84
4.1.1.4. Przegląd badań naukowych	85
4.1.2. Opis eksperymentu	88
4.1.2.1. Cel badania	88
4.1.2.2. Dostępne dane empiryczne	89
4.1.2.3. Problem hakowania systemów uczących się	91
4.1.3. Wnioski i rekomendacje	105
Bibliografia	107
4.2. Atak na system detekcji nadużyć w bankowości elektronicznej	110
4.2.1. Problem nadużyć w bankowości elektronicznej	110
4.2.1.1. Wprowadzenie	110
4.2.1.2. Definicja nadużycia w transakcjach bankowych	111
4.2.1.3. Wykrywanie nadużyć i przeciwdziałanie im	112
4.2.1.4. Standardowy system wykrywania i przeciwdziałania nadużyciom	113
4.2.2. Opis eksperymentu	115
4.2.2.1. Cel badania	115
4.2.2.2. Dostępne dane empiryczne	117
4.2.2.3. Generatywne sieci współzawodniczące (GANs)	118
4.2.2.4. Scenariusze przebiegu ataku	122
4.2.3. Modele generatora i dyskryminatora	125
4.2.3.1. Budowa modeli	125
4.2.3.2. Ewaluacja modeli	129
4.2.4. Wnioski końcowe i rekomendacje	132
Bibliografia	133
5. Bezpieczeństwo aplikacji systemów uczących się	137
5.1. Wprowadzenie	138

5.2. Wybrane problemy niezawodności oprogramowania	139
5.2.1. Problem złożoności kodu	139
5.2.2. Przepiętnienie bufora oraz odczyt poza nim	141
5.2.3. Dostęp po zwolnieniu pamięci	142
5.2.4. Niewłaściwa deserializacja i wstrzykiwanie danych	142
5.3. Ataki na środowiska programistyczne i sprzęt dla systemów uczących się	143
5.3.1. Atak na platformę programistyczną	143
5.3.2. Atak na sprzęt na przykładzie Deep Hammer	145
5.4. Ataki na biblioteki z wykorzystaniem automatycznych metod testowania oprogramowania	146
5.4.1. Wprowadzenie	146
5.4.2. Atak na bibliotekę OpenCV	147
5.4.3. Atak na bibliotekę dlib	150
5.4.4. Podsumowanie podatności znalezionych za pomocą automatycznych metod testowania oprogramowania	152
5.5. Wnioski i kierunki dalszego działania	153
Bibliografia	154
Zakończenie	157
Bibliografia	159

oprac. BPK