

FAIK

SZTUCZNA INTELIGENCJA
W SŁUŻBIE FAŁSZYWEJ
RZECZYWISTOŚCI

JAK PRZETRWAĆ W EPOCE
CYFROWYCH OSZUSTW

PERRY CARPENTER

Tytuł oryginału: FAIK: A Practical Guide to Living in a World of Deepfakes,
Disinformation, and AI-Generated Deceptions

Tłumaczenie: Zbigniew Waśko

ISBN: 978-83-289-3107-7

Copyright © 2025 by John Wiley & Sons, Inc. All rights, including for text
and data mining, AI training, and similar technologies, are reserved.

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

Translation © 2026 by Helion S.A.

No part of this publication may be reproduced, stored in a retrieval system,
or transmitted in any form or by any means, electronic, mechanical, photocopying,
recording, scanning, or otherwise, without the prior written permission of the Publisher.

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.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości
lub fragmentu niniejszej publikacji w jakiegokolwiek 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/faiksz

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

Przedmowa	9
Wprowadzenie. Odpowiedz mi, bocie ChatGPT	13
Rozdział 1. Odwieczna walka o umysł — dlaczego powinno Cię to obchodzić?	25
Szepty w szumie	25
Miło Cię poznać, mam nadzieję, że odgadłeś moje imię	27
Oszustwa i przekręty w kontekście historycznym	29
Era cyfrowa — nowy wiek oszustwa	31
Dalekosiężne konsekwencje istnienia mediów syntetycznych	38
Dlaczego syntetyczne media omijają nasze zabezpieczenia?	39
Podstawy bezpieczeństwa poznawczego — ochrona umysłu w erze cyfrowej	40
Podsumowanie	43
Rozdział 2. Nowe granice oszustwa — sztuczna inteligencja i media syntetyczne	45
Szepty w szumie	45
Punkty przełomowe	47
Krótką historią rozwoju sztucznej inteligencji	47
Czy sztuczna inteligencja jest żywa i czy naprawdę rozumie?	61
Podsumowanie	67

Rozdział 3.	Sposób myślenia i narzędzia cyfrowego manipulatora	69
	Szepty w szumie	69
	Jak hakerzy podchodzą do techniki i rozwiązywania problemów?	72
	Podstawy deceptionsologii — wprowadzenie do mrocznych sztuk	77
	Inspekcja króliczej nory	89
	Podsumowanie	94
Rozdział 4.	Stronniczość, zatrutowanie danych i osobliwości wyników	97
	Szepty w szumie	97
	Halucynacje — obosieczny miecz kreatywności sztucznej inteligencji	99
	Duży problem uprzedzeń	104
	Wstydlive porażki w kontrolowaniu stronniczości AI	107
	Konsekwencje i rozwiązanie	118
	Podsumowanie	119
Rozdział 5.	Pandemia cyfrowej dezinformacji	121
	Szepty w szumie	121
	Kłamstwa, cholerne kłamstwa i internet	123
	Krajobraz dezinformacji napędzanej sztuczną inteligencją	132
	Rodzaje i taktyki dezinformacji generowanej przez sztuczną inteligencję	134
	Podsumowanie	140
Rozdział 6.	Deepfaki i spektrum cyfrowego oszustwa	142
	Szepty w szumie	142
	W internecie nikt nie wie, że jesteś psem	144
	Ale czym jest ten cały deepfake?	146
	Zatrważająca skuteczność prostych metod oszustwa	151
	Breaking Bad — jak oszuści mogą zepsuć moralność systemów AI?	158
	Gdy jailbreaking jest niepotrzebny, czyli nieocenzurowane systemy AI	178
	Uzbrajanie niewinnych odpowiedzi	182
	Nieprzyjemna prawda o cyfrowych oszustwach	186
	Podsumowanie	187
Rozdział 7.	Teraźniejszość i przyszłość oszustw napędzanych sztuczną inteligencją	190
	Szepty w szumie	190
	W głąb strefy eksploatacji	193
	Filary oszustwa, przekrętów i przestępczości	196
	Oszustwa wspomagane sztuczną inteligencją — stare sztuczki, nowe narzędzia	200

Nowe zagrożenia	207
Historia osobista — jak stworzyłem wiele oszukańczych botów wykorzystujących generatywną AI?	215
W poszukiwaniu nadziei	218
Podsumowanie	218
Rozdział 8. Umiejętne korzystanie z mediów w erze AI — pierwsza linia obrony	221
Szepty w szumie	221
Walka o prawdę	224
Krótki wstęp przed rozpoczęciem	226
Kompetencje medialne w erze cyfrowej — przewodnik po dżungli informacyjnej	231
Paradoks weryfikacji faktów — poszukiwanie prawdy w świecie kłamstw	240
Podsumowanie	254
Rozdział 9. Strategie obrony kognitywnej i technicznej — nasze środki ochrony	258
Szepty w szumie	258
Budowanie obrony poznawczej	261
Obrona oparta na technologii	278
W skrócie: era cyfrowa wymaga czujności	289
Do dzieła: rozwijaj swoje umiejętności obrony mentalnej i cyfrowej	290
Podsumowanie	293
Rozdział 10. Nowa nadzieja	296
Szepty w szumie	296
Przestępcy mają środki, motyw i okazję. My również	299
Do dzieła — niech się dzieje	305
Podsumowanie	308
 Dodatek. Wskazówki, triki i zasoby	 315
Przypisy	326
Podziękowania	340
O autorze	343

Rozdział 1.

Odwieczna walka o umysł — dlaczego powinno Cię to obchodzić?

Szepty w szumie

To był wyczerpujący dzień. Leżysz na kanapie, próbując się odprężyć, bezmyślnie przeglądając Facebooka i lajkując filmiki z kotami oraz zdjęcia z wakacji. Nagle jeden post przykuwa twoją uwagę. To twoja najlepsza przyjaciółka, Sara. Płacze, jest zrozpaczona.

— Nie mogę w to uwierzyć — wykrztusza z trudem Sara. Ma zaczerwienioną twarz i ochrypły głos. — Moja mama... odeszła. Zmarła w nocy. To było tak nagłe, wszyscy jesteśmy w szoku.

Czujesz, jak żołądek podchodzi ci do gardła. Mama Sary? Nie żyje? To niemożliwe.

Sara bierze urywany oddech, po czym patrzy prosto w kamerę i mówi:

— Założyliśmy zbiórkę na GoFundMe, żeby zdobyć pieniądze na pogrzeb. Wszystko dzieje się tak szybko, a koszty są... Każda pomoc będzie dla nas bezcenna. Naprawdę mamy teraz ciężko.

Sięgasz po telefon, żeby zadzwonić. Ale po chwili namysłu rezygnujesz, bo zdajesz sobie sprawę, że Sara pewnie ma teraz mnóstwo spraw na głowie. Nie chcesz dokładać jej stresu i zamieszania, więc postanawiasz wysłać krótką wiadomość: „Boże, nie mogę w to uwierzyć. Tak mi przykro”.

Godzinę później twój telefon wibruje. To Sara. Przygotowujesz się psychicznie i odbierasz.

— Hej, widziałam twój filmik. Tak mi przykro z powodu twojej mamy.

— Co? Z moją mamą wszystko w porządku, właśnie jadłam z nią kolację. Dzwonię, bo chciałam się dowiedzieć, o co ci chodziło w tej wiadomości. Z jakiego powodu jest ci przykro? O jakim filmiku mówisz?

Zamierasz ze zdziwienia.

— O filmiku... na Facebooku. Płakałaś, mówiłaś, że twoja mama zmarła i prosiłaś o datki.

Cisza. A po chwili:

— To nie ma sensu. Wyślij mi ten filmik. Natychmiast.

Drżącymi palcami wysyłasz jej link.

— To nie jestem ja — mówi Sarah po dłuższej chwili, głosem drżącym ze strachu, dezorientowania i gniewu jednocześnie. — Nie rozumiem... Ktoś sfalszował to nagranie. To moja twarz, mój głos. I mój pokój w tle. Ale to nie jestem ja. Nie nagrałam tego.

Świadomość konsekwencji osiada ci ciężko w żołądku. Jeśli można tak realistycznie sfalszować wideo... Jeśli tak łatwo oszukało ciebie, najlepszą przyjaciółkę Sary... To co jeszcze można podrobić? Jakie inne kłamstwa można rozpowszechniać jednym kliknięciem?

— O Boże, Sara. Ludzie będą wpłacać pieniądze. Uwierzą, że to prawda.

Sarah przeklina pod nosem:

— Muszę to zgłosić, żeby usunęli tę zbiórkę. Nie mogę uwierzyć, że ktoś mógł to zrobić. Co by było, gdyby to zobaczyła moja babcia albo młody kuzynek?

Oniemiała, możesz odpowiedzieć tylko milczeniem. Milczeniem, któremu towarzyszy ponura świadomość, że to nowa rze-

czywistość. Rzeczywistość, w której zaufanie może z przerażającą łatwością być wykorzystane jako broń. To świat, na który nie jesteś gotowa, ale to jedyny świat, jaki masz. I gdy razem z Sarą szukacie sposobu, by zgłosić fałszywe wideo i stronę zbiórki, wiesz, że od tej pory nie możesz niczego brać za pewnik. Nawet lez swojej najlepszej przyjaciółki.

Miło Cię poznać, mam nadzieję,
że odgadłeś moje imię

Jeśli istnieje coś fundamentalnego dla ludzkiej natury, to jest to oszustwo. Wystarczy się nad tym chwilę zastanowić: motywy związane z kłamstwem i oszustwem przenikają wszystko — od baśni, w których wielki zły wilk przebiera się za kruchą babcię, po ulubione filmy szpiegowskie, w których agenci przywdziewają stroje ekipy sprzątającej, by uniknąć wykrycia. Te historie o oszustwie zapadają nam w pamięć, ponieważ odzwierciedlają nasze doświadczenia. Być człowiekiem to oszukiwać i być oszukiwanym.

Jednak z każdą epoką i każdym postępem oszustwo ewoluuje. Jest to szczególnie widoczne w naszej obecnej erze cyfrowej. I ten proces przyspiesza. Ostatnie postępy w dziedzinie **sztucznej inteligencji** (AI), a zwłaszcza **generatywnej sztucznej inteligencji** (GenAI), dały każdemu możliwość tworzenia wiarygodnych oszustw na wielką skalę i w formach praktycznie nie do odróżnienia od rzeczywistości.

Bez wątpienia technologia osiągnęła punkt, w którym naszym największym zmartwieniem nie są już fałszywe wiadomości. *Fałszywe może być wszystko.* Tworzenie fałszywych rzeczywistości zostało zdemokratyzowane.

Władza dla ludu... władza przekonującego oszukiwania, rzecz jasna.

Dzisiejsze kłamstwa są „turbodoładowane” przez technologię. I przez to trudno je wykryć. Deepfaki mogą sprawić, że światowi przywódcy wydają się mówić rzeczy, których nigdy nie powiedzieli, pokazywać celebrytów w kompromitujących sytuacjach, które nigdy nie miały miejsca, i skłonić nas do kwestionowania samej natury tego, co jest rzeczywiste. **Grafika generowana komputerowo** (CGI) może wyczarować wydarzenia z niczego, wraz z fałszywymi materiałami informacyjnymi, które wyglądają aż nazbyt prawdziwie. Armie botów mogą zalać nasze media społecznościowe sztucznie wykreowanym oburzeniem. Rezultat? Możemy zostać przekonani, że marginalne idee nagle stały się głównym nurtem.

Witajcie w odważnym nowym świecie deepfake’ów: świecie, w którym wpływowe postacie stają się nieświadomymi marionetkami, a ich cyfrowe wizerunki zostają przejęte, by szerzyć sfa-brykowane prawdy. Sama natura rzeczywistości staje się przedmiotem manipulacji, a odróżnienie prawdy od przekonującej iluzji staje się coraz trudniejsze. W tej nowej cyfrowej dystopii, gdzie piksele mogą kłamać równie łatwo jak słowa, jak nauczyć się oddzielać fakty od fikcji? Czy zmierzamy w stronę świata, w którym nawet pojęcia *prawda* i *fakty* są zagrożone?

Aby mieć szansę w starciu z zalewem tego nowego fałszu, musimy najpierw zrozumieć ponadczasowe zasady stojące za skutecznością oszustw. W tym celu udajmy się w podróż przez barwną przeszłość dawnych oszustów. Śledząc ewolucję oszustwa od starożytnych mitów po współczesne memy, odkryjemy trwałe ludzkie dziwactwa i kognitywne skłonności, które dzisiejsi technoszuści tak umiejętnie wykorzystują.

Oszustwa i przekręty w kontekście historycznym

Przez tysiąclecia, od starożytnych wodzów prezentujących fałszywe relikwie po współczesnych prezesów fałszujących księgi rachunkowe, oszuści doskonalili swoje metody. Technologia może się zmieniać, ale podstawowe taktyki pozostają te same: wykorzystaj zaufanie, wykorzystaj chciwość i użyj strachu jako broni. Nazwiska i twarze zmieniają się wraz z upływem wieków, ale istota gry jest stara jak ludzka natura.

Starożytne początki: od konia trojańskiego do sprzedawców cudownych eliksirów

Karty historii są pełne opowieści o oszustwach i oszustach. Weźmy słynnego konia trojańskiego, prawdopodobnie najbardziej legendarny przykład podstępów. Pomyślcie tylko: ten przykład jest tak ikoniczny, że ludzie używają terminów **podstęp** i **koń trojański** niemal zamiennie. Jednym ogromnym aktem pozornej życzliwości Grecy oszukali swoich wrogów i sprawili, że ci z radością sprowadzili na siebie zgubę. To dobitne przypomnienie, że czasami najniebezpieczniejsze kłamstwa to te, w które tak desperacko chcemy wierzyć.

Albo weźmy wyrocznie delficką, najbardziej znane w starożytnym świecie źródło przepowiedni. Dzięki swoim słynnym z niejasności oświadczeniom, które można było interpretować na wiele sposobów, wyrocznia stworzyła dochodowy interes. Mówiła ludziom dokładnie to, co chcieli usłyszeć. Wykreowała schemat, który współcześni jasnowidze, astrologi, specjaliści od marketingu i komentatorzy telewizyjni bezwstydnie kopiują do dziś¹.

Przenieśmy się o kilka tysięcy lat do przodu, do czasów sprzedawców cudownych eliksirów — tych krasomówczych handlarzy mikstur i leków na wszystko: od fałszywych suplementów na

odchudzanie po oszukańcze toniki na porost włosów. Ci oszuści zbijali fortuny, żerując na odwiecznym ludzkim pragnieniu szybkiego i łatwego rozwiązywania problemów. Nie miało znaczenia, że ich mikstury często były niczym więcej niż aromatyzowaną wodą lub alkoholem. Prawdziwym magicznym składnikiem zawsze była sztuka perswazji.

Ewolucja oszustw: dostosowywanie się do nowych technologii

Wraz z rozwojem narzędzi ewoluowały także taktyki oszustów. Z pojawieniem się prasy drukarskiej nastąpił rozkwit podrabianych dokumentów, od fałszywych pieniędzy po sfalszowane akty własności. Telegraf zapoczątkował erę oszustw na odległość, w których manipulanci podszywali się pod dalekich krewnych w rzekomej potrzebie, aby wyłudzić od ofiar pieniądze. A wraz z nastaniem ery cyfrowej wrota oszustwa naprawdę się otworzyły.

Ale to pojawienie się mediów masowych wyniosło sztukę mistyfikacji na zupełnie nowy poziom. Panika wywołana słuchowiskiem radiowym Orsona Wellesa *Wojna światów* z 1938 roku — gdy miliony ludzi uwierzyły, że Marsjanie naprawdę atakują New Jersey — zapoczątkowała nową erę, w której kłamstwa mogły być wzmacniane i rozpowszechniane jak nigdy dotąd². Jak się okazało, ta „masowa panika” była również kłamstwem, wymyślonym przez przebiegłą gazetę, chcącą zdyskredytować rozwijające się medium radiowe³. Co ciekawe, w obu przypadkach mamy do czynienia z oszustwem. W obu widzimy, jak łatwo można na masową skalę manipulować rozumieniem rzeczywistości.

Era cyfrowa — nowy wiek oszustwa

Z wielką technologią przychodzą wielkie możliwości.

Czy to prasa drukarska, fale radiowe czy niemal świetlna prędkość internetu* — oszustwo rozprzestrzenia się z prędkością, na jaką pozwala technologia. Żyjemy w czasach, gdy armie botów i farmy trolli są gotowe w każdej chwili zasypać media społecznościowe dezinformacją. Deepfaki i inne **syntetyczne media**[†] mogą fabrykować dowody audiowizualne z niczego. W tej nowej rzeczywistości, gdzie każdy głos może zostać sklonowany, a każda twarz podmieniona, stare powiedzenie „zobaczyć znaczy uwierzyć” przestaje mieć zastosowanie.

Rozwój internetu: rozszerzanie zasięgu oszustów

Internet zrewolucjonizował sposób działania oszustów. Nie są już ograniczeni do bezpośrednich kontaktów czy zaufkowych transakcji — teraz jednym kliknięciem mogą dotrzeć do milionów potencjalnych ofiar na całym świecie. Anonimowość cyfrowego świata ułatwia oszustwa i utrudnia ich wykrycie. A niemal natychmiastowa prędkość globalnej komunikacji oznacza, że cyfrowi oszuści mogą sięgać dalej i szybciej niż kiedykolwiek wcześniej.

Oto przykład, jak rozwój internetu może wspomóc ewolucję oszustwa. Zapewne słyszałeś o słynnym spamowym „szwindlu nigeryjskim”. To klasyczne oszustwo znane jako **wyłudzenie**

* Tak... No cóż. Mój domowy internet jeszcze o tym nie wie.

[†] **Syntetyczne media** to takie fantazyjne określenie treści wygenerowanych przez sztuczną inteligencję. Mamy z tym do czynienia, gdy algorytmy komputerowe tworzą, modyfikują lub mieszają informacje oraz ich nośniki takie jak wideo, obrazy i nagrania audio. Klasycznym tego przykładem są deepfaki — treści zmanipulowane przez AI, często by oszukać ludzi lub zmienić pierwotny sens informacji.

zaliczki. Zaczyna się od e-maila, w którym ktoś podający się za nigeryjskiego arystokratę lub wysokiego urzędnika kontaktuje się z ofiarą, aby złożyć jej niesamowitą propozycję. Opowiada historię o dużej sumie pieniędzy, do której nie ma dostępu, i obiecuje znaczny udział w zamian za pomoc. Wystarczy wpłacić zaliczkę na pokrycie różnych fikcyjnych kosztów. Gdy ofiara zapłaci, gra się kończy. Oszust znika z pieniędzmi, a ofiara zostaje z niczym.

Ale jak mówi przysłowie, „co stare, to nowe”. Korzenie szwindlu nigeryjskiego sięgają znacznie dalej w przeszłość, niż mogłoby się wydawać. Można powiedzieć, że Nigeryjczyk to tak naprawdę Hiszpan w przebraniu. To ewolucja znacznie starszego przekrętu znanego jako oszustwo „na hiszpańskiego więźnia”⁴. W oryginalnej wersji, która pojawiła się pod koniec XIX wieku, oszust wysyłał listy, w których podawał się za bogatego więźnia potrzebującego pomocy w dostępie do swojej fortuny i obiecywał udział w bogactwie w zamian za pomoc. Oczywiście „pomoc” polegała na wpłaceniu pieniędzy. Pieniądzy na pokrycie różnych opłat, łapówek czy wydatków, aby móc odzyskać fortunę. Uwaga... żadnej fortuny nie było. Po drugiej stronie był tylko oszust, który właśnie wzbogacił się o pieniądze ofiary. W erze cyfrowej to samo oszustwo rozprzestrzeniło się przez e-maile, docierając do znacznie szerszego grona odbiorców.

To tylko jeden przykład. Oszustwa rozwijają się w tempie postępu technologicznego. Wraz z rozwojem technologii ewoluowały też przekręty. Popularną strategią stał się phishing polegający na wysyłaniu e-maili starannie spreparowanych, tak by przypominały wiadomości od legalnych firm, i nakłaniających odbiorców do ujawnienia poufnych informacji. Pojawiły się fałszywe strony internetowe, gotowe wykraść dane kart kredytowych pod pretekstem niesamowitych okazji. Internet stworzył — i nadal tworzy — żyzny grunt dla tych, którzy chcą doskonalić swoje techniki oszukiwania i zarzucać coraz szerszą sieć.

*Nadejście mediów społecznościowych —
idealnej platformy dla dezinformacji*

Media społecznościowe, początkowo chwalone jako narzędzie do budowania relacji i społeczności, przekształciły się w wylęgarnię dezinformacji. Wśród miliardów aktywnych użytkowników na platformach takich jak Facebook, X (dawniej Twitter) i Instagram fałszywe informacje mogą stać się wiralem w ciągu kilku godzin⁵. Co więcej, rozprzestrzeniają się „dalej, szybciej, głębiej i szerzej niż prawda”⁶. Im bardziej szokująca i emocjonalna jest treść, tym większe prawdopodobieństwo, że zostanie szeroko udostępniona⁷.

To zjawisko napędziło wzrost **fake newsów** — historii celowo stworzonych, aby wprowadzać w błąd i manipulować. Żli ludzie rozumieją ogromną moc mediów społecznościowych w kształtowaniu opinii publicznej i wykorzystują je do niecznych celów. Od wpływania na wybory po sianie niezgody społecznej; wpływ kampanii dezinformacyjnych może być daleko idący i niszczy-cielski.

Ale nie chodzi tylko o treść. Algorytmy mediów społecznościowych, zaprogramowane tak, aby utrzymać uwagę użytkowników, często tworzą **bańki informacyjne**⁸ — przestrzenie, w których użytkownicy otrzymują głównie informacje zgodne z ich przekonaniami i zainteresowaniami. Jak można sobie wyobrazić, ma to efekt wzmacniania uprzedzeń i zacierania granicy między faktem a fikcją. Ten błędny cykl przyczynia się do zwiększonej polaryzacji i podważa podstawy konstruktywnego dialogu.

Punkt zwrotny sztucznej inteligencji — automatyzacja i udoskonalenie oszustwa

Jakby krajobraz nie był już wystarczająco złożony, rozwój sztucznej inteligencji dodaje kolejną warstwę do problemu. Dziedziny takie jak uczenie maszynowe, przetwarzanie języka naturalnego i generatywna AI ułatwiają automatyzację i zwiększanie skali oszustw na bezprecedensową skalę.

Deepfaki, na przykład, to przekazy generowane przez sztuczną inteligencję, które mogą przedstawiać kogokolwiek mówiącego lub robiącego cokolwiek, niezależnie od tego, czy taka sytuacja faktycznie miała miejsce. Potencjał nadużyć jest ogromny, od sabotażu politycznego po osobiste porachunki.

Definicja deepfake'u

Termin **deepfake** nie zawsze jest właściwie rozumiany. Ogólnie, jest to połączenie dwóch różnych terminów: **głębokoci** (ang. *deep*) i **fałszywy** (ang. *fake*). Deep odnosi się tu do stosowania procesu **uczenia głębokiego** (ang. *deep learning*). Nie musimy się zagłębiać zbytnio w ten proces. Wystarczy wiedzieć, że słowo deep określa tu, jak dobrze sieć neuronowa jest w stanie odtworzyć bardzo realistyczny przekaz na podstawie danych treningowych.

A *fake* oznacza, no cóż... wiadomo... fałszywy. Po prostu fałszywy.

Deepfake to niezwykle realistyczny i zwodniczy materiał wygenerowany przy użyciu technik głębokiego uczenia. Choć termin ten może odnosić się do dowolnych syntetycznych mediów stworzonych przez sztuczną inteligencję, większość ludzi używa go do określenia złośliwego wykorzystania takich materiałów. Co więcej, niektó-

rzy stosują to pojęcie w odniesieniu do wszelkich wprowadzających w błąd treści stworzonych przy pomocy technologii, niezależnie od tego, czy w proces ten zaangażowano sztuczną inteligencję.

Zjawisko to nie ogranicza się tylko do nagrań wideo. Sztuczna inteligencja może również z zatrważającą wydajnością generować fałszywe teksty, tworząc przekonujące artykuły, posty w mediach społecznościowych, a nawet całe strony internetowe pełne dezinformacji — a wszystko to w tempie i skali niemożliwych do osiągnięcia przez człowieka. Szczególnie niepokojący jest fakt, że spersonalizowany chatbot AI, dysponujący nawet niewielką ilością informacji o Twoich zainteresowaniach i przekonaniach, często skuteczniej zmienia Twoje zdanie niż żywa osoba⁹. Innymi słowy, im więcej model AI wie o Tobie, tym lepiej może dostosować swoje kłamstwa, aby wykorzystać Twoje unikalne uprzedzenia i słabości.

W miarę rozwoju sztucznej inteligencji te zwodnicze techniki będą stawać się coraz bardziej wyrafinowane i trudniejsze do wykrycia. Stoimy przed poważnym wyzwaniem w walce z cyfrowym oszustwem we wszystkich jego wciąż ewoluujących formach.

Połączenie internetu, mediów społecznościowych i sztucznej inteligencji stworzyło podatny grunt dla rozkwitu oszustw. To trudna rzeczywistość, z którą musimy się zmierzyć. Konsekwencje tej nowej ery wykraczają daleko poza sferę cyfrową, wpływając na jednostki, firmy i całe społeczeństwa.

Rozwój deepfake'ów — od celebryckich twarzy do sabotażu politycznego

W kwietniu 2017 roku artysta Cameron-James Wilson wywołał na Instagramie poruszenie serią zdjęć oszałamiającej modelki Shudu Gram^{10,11}. Shudu, chwalona za swoją wyjątkową urodę, szybko zdobyła ogromną rzeszę fanów. Był jednak haczyk: Shudu nie była prawdziwa. Była hiperrealistyczną trójwymiarową cyfrową kreacją wymyśloną przez Wilsona.

W ciągu ostatnich kilku lat obrazy i filmy generowane przez sztuczną inteligencję stały się jeszcze bardziej wyrafinowane i doprowadziły do wielu przypadków, w których opinia publiczna dała się zwieść ich realizmowi. Oto kilka przykładów:

- **Papież w puchowej kurtce.** Wygenerowany przez AI obraz papieża Franciszka ubranego w modną białą kurtkę puchową¹² stał się wiralem w mediach społecznościowych. Wiele osób, włącznie z celebrytami takimi jak Chrissy Teigen, uwierzyło, że zdjęcie jest prawdziwe¹³.
- **Aresztowanie Donalda Trumpa.** Wygenerowane przez AI obrazy przedstawiające byłego prezydenta Donalda Trumpa podczas aresztowania wywołały poruszenie w mediach społecznościowych. Mimo szerokiego rozpowszechnienia, nie były jednak prawdziwe i stanowiły część narracji stworzonej przy użyciu Midjourney, generatora obrazów AI¹⁴.
- **Dzieła generowane przez AI wygrywają konkursy.** Dzieło zatytułowane *Théâtre D'opéra Spatial*, stworzone przez Jasona Allena przy użyciu programu Midjourney, zdobyło pierwsze miejsce w konkursie Colorado State Fair. Wywołało to debatę na temat roli AI w sztuce i natury kreatywności¹⁵.
- **Fałszywe zdjęcia historyczne.** AI została wykorzystana do tworzenia fałszywych zdjęć historycznych, które wyglądają autentycznie¹⁶, na przykład obraz przedstawiający Elona

Muska jako bezdomnego z lat 20. poprzedniego wieku¹⁷. Te obrazy mogą być szczególnie mylące, ponieważ płynnie wtapiają się w autentyczne zapisy historyczne. Wyobraźmy sobie, co może wynikać z przekonującego przepisania historii w połączeniu z sugestywnymi zdjęciami.

- **Deepfake’owe nagrania celebrytów.** Za pomocą technologii deepfake’owej stworzono filmy z udziałem celebrytów takich jak Taylor Swift i Tom Hanks, a następnie wykorzystano te filmy w oszustwach bez pytania bohaterów o zgodę¹⁸. Filmy były na tyle przekonujące, że wprowadziły w błąd fanów i były szeroko rozpowszechniane w mediach społecznościowych.
- **Fałszywe oskarżenie.** W styczniu 2024 roku w mediach społecznościowych pojawiło się szokujące 42-sekundowe nagranie głosowe. Klip, rzekomo nagrany potajemnie, przedstawiał dyrektora liceum w Maryland wygłaszającego rasistowskie komentarze. W wyniku oburzenia publicznego dyrektor został zawieszony w obowiązkach. Później okazało się jednak, że nagranie było deepfakiem stworzonym przez niezadowolonego pracownika po sporze dotyczącym umowy¹⁹.
- **Oszustwo z udziałem dyrektora finansowego.** W Hongkongu pracownik działu finansowego został oszukany i wypłacił 25 milionów dolarów po uczestnictwie w videokonferencji z deepfake’owymi reprezentacjami pracowników firmy, w tym dyrektora finansowego. W rzeczywistości pracownik finansowy był jedyną prawdziwą osobą na rozmowie. Deepfaki były na tyle przekonujące, że przezwyciężyły początkowe podejrzenia pracownika²⁰.

To wszystko świadczy o rosnących trudnościach w odróżnianiu prawdziwych treści od tych generowanych przez sztuczną inteligencję oraz o potencjale wykorzystania tej technologii do celowego wprowadzania w błąd. Przez wieki polegaliśmy na aksjomacie „zobaczyć znaczy uwierzyć”. Jednak w erze mediów

syntetycznych jest on nieaktualny. Treści generowane przez AI potrafią fabrykować realia nieodróżnialne od prawdziwych.

Konsekwencje są ogromne. Jeśli każde zdjęcie, film czy nagranie dźwiękowe może być fałszywe, skąd mamy wiedzieć, którym można zaufać? Jak poruszać się w rzeczywistości, w której nasze zmysły mogą być oszukiwane na każdym kroku?

Dalekosiężne konsekwencje istnienia mediów syntetycznych

Technologia tworzenia niezwykle realistycznych fałszywych treści rozwija się w zawrotnym tempie. Narzędzia do tworzenia podstawowych deepfake'ów istnieją od lat, w różnym stopniu zaawansowania i wiarygodności. Jednak w momencie, gdy piszę te słowa w 2024 roku, osiągnęliśmy punkt, w którym syntetyczne media są praktycznie — a często całkowicie — nie do odróżnienia od rzeczywistości²¹.

Potencjalne zastosowania są alarmujące. Wyobraźmy sobie ucznia szkoły średniej wykorzystującego technologię deepfake do tworzenia kompromitujących nagrań swojego nauczyciela. Albo niezadowolonego pracownika używającego klonowania głosu opartego na sztucznej inteligencji do generowania fałszywych nagrań, na których jego szef wypowiada dyskryminujące uwagi. Możliwości nękania, zastraszania i niszczenia reputacji są nieograniczone.

Na poziomie społecznym implikacje są jeszcze bardziej oszałamiające. Sfabrykowane nagrania polityków przyjmujących łapówki, fałszywe klipy audio prezesów firm ujawniających poufne informacje, zmyślone artykuły cytujące lekarzy zalecających niebezpieczne cudowne leki — wszystko to może mieć dalekosiężne konsekwencje. Gdy wszystko staje się nasyczone syntetycznymi mediami, naszego postrzeganie rzeczywistości zaczyna się rozpadać.

Dlaczego syntetyczne media omijają nasze zabezpieczenia?

Oto sekret, który wcale nie jest tajny: jesteś łatwowierny. Ja też taki jestem. Wszyscy tacy jesteście. To nie obelga, to po prostu ludzka natura. Nasze mózgi są zaprogramowane, by ufać temu, co widzą nasze oczy i słyszą nasze uszy. Z natury jesteście podatni na oszustwa. A im bardziej realistyczne stają się syntetyczne media, tym bardziej nasz wrodzony „wewnętrzny sceptyk” staje się bezradny.

Pomyśl o swoich ostatnich miesiącach w internecie. Założę się, że pamiętasz przynajmniej jeden przypadek, gdy natknąłeś się na post, film lub inną formę treści, która przykuła Twoją uwagę. Może to było nagranie polityka lub celebryty mówiącego coś szokującego lub nietypowego. I może później dowiedziałeś się, że to nie było prawdziwe. Ale bardzo prawdopodobne, że Twoją pierwszą myślą nie było: „To musi być fałsz”. Zamiast tego prawdopodobnie pomyślałeś: „Wow, nie mogę uwierzyć, że to powiedzieli!”.

Wszyscy mamy naturalną skłonność do ufania temu, co wydaje się autentycznym dźwiękiem i obrazem. To ten sam powód, dla którego zeznania naocznych świadków są tak przekonujące w sądzie, mimo że mnóstwo badań pokazuje, jak zawodna jest ludzka pamięć²². Po prostu jest w nas coś, co każe nam wierzyć ludzkim zmysłom.

Złoczyńcy tworzący wprowadzające w błąd syntetyczne media rozumieją to i wykorzystują. Ale też idą o krok dalej, wykorzystując inną kluczową ludzką słabość: emocje. Jak już wspomnieliśmy w części o platformach społecznościowych, badania wielokrotnie pokazywały, że treści wywołujące intensywną reakcję emocjonalną — czy to gniew, strach, obrzydzenie czy radość — są znacznie łatwiej zapamiętywane i chętniej przekazywane.

Dlatego twórcy oszukańczych syntetycznych mediów projektują je tak, by naciskać nasze emocjonalne guziki, co praktycznie gwarantuje, że te zniekształcone informacje staną się wiralami. Wykorzystując zarówno nasze wrodzone zaufanie do treści audiowizualnych, jak i naszą skłonność do reagowania emocjonalnie zamiast racjonalnie, mogą ominąć nasze psychologiczne mechanizmy obronne i zmanipulować nas, byśmy uwierzyli w fałszywe informacje i je rozpowszechniali.

To podstępny cios. A w miarę postępu technologii te zmanipulowane media będą coraz bardziej naturalne i przekonujące. Bez dalszego treningu naszego wewnętrznego sceptyka po prostu nie będziemy w stanie odróżnić prawdziwego wideo od dobrze wykonanego deepfake'u.

Co możemy więc zrobić? Jak mamy bronić nasze umysły przed kłamstwami, które wyglądają, brzmią i wydają się prawdziwe? Kluczem jest rozwijanie naszych kognitywnych mechanizmów obronnych.

Podstawy bezpieczeństwa poznawczego — ochrona umysłu w erze cyfrowej

Bezpieczeństwo poznawcze polega na ochronie naszej zdolności do trzeźwego myślenia, logicznego rozumowania i podejmowania dobrych decyzji w obliczu coraz bardziej wyrafinowanych prób wpływania na nas. Tak jak cyberbezpieczeństwo chroni nasze systemy cyfrowe, tak bezpieczeństwo poznawcze ma na celu ochronę naszych procesów myślowych.

Nasze zdolności poznawcze — umiejętność zdobywania wiedzy, postrzegania rzeczywistości i wydawania sądów — są fundamentalne dla tego, kim jesteśmy i jak radzimy sobie w życiu. W erze cyfrowej złoźnicy mają bezprecedensowe możliwości wykorzystywania ludzkich słabości poznawczych przez szybkie

rozprzestrzenianie błędnych informacji w mediach społecznościowych i za pośrednictwem innych platform internetowych.

Deepfaki, teorie spiskowe i podburzająca retoryka mogą się rozprzestrzeniać w sieci jak pożar, przyciągając naszą uwagę i manipulując emocjami. Długotrwała ekspozycja na tego typu treści może osłabiać naszą zdolność do odróżniania faktów od fikcji, kształtowania wyważonych opinii i podejmowania rozsądnych decyzji — pogarszając nasze indywidualne samopoczucie i zbiorową zdolność do funkcjonowania jako społeczeństwo.

Tu właśnie wkracza bezpieczeństwo poznawcze. Czerpiąc z nauk behawioralnych, neurobiologii i informatyki, badacze z tej rozwijającej się dziedziny pracują nad zrozumieniem, w jaki sposób nasze umysły przetwarzają informacje i jak można nimi manipulować. Następnie wykorzystują tę wiedzę do opracowywania strategii i technik wzmacniania naszych poznawczych mechanizmów obronnych.

Obejmuje to między innymi:

- badanie wpływu różnych rodzajów mediów na procesy poznawcze oraz opracowywanie modeli mentalnych i technik krytycznej oceny informacji,
- tworzenie programów edukacyjnych rozwijających odporność poznawczą, umiejętność korzystania z mediów i krytyczne myślenie od najmłodszych lat,
- projektowanie opartych na sztucznej inteligencji narzędzi do wykrywania i zwalczania zmanipulowanych treści, kampanii informacyjnych z udziałem botów i innych form ataków poznawczych przeprowadzanych w tempie i skali internetu,
- integrowanie zbiorów danych i wiedzy z różnych dyscyplin w celu lepszego zrozumienia i przeciwdziałania ewoluującym zagrożeniom dla bezpieczeństwa poznawczego.

Odporność poznawcza wymaga od nas wykształcenia nawyków krytycznego myślenia i świadomego korzystania z mediów. Oznacza to naukę rozpoznawania charakterystycznych oznak dezinformacji, takich jak manipulacja emocjonalna, fałszywy kontekst i niesprawdzone twierdzenia. Wymaga to poświęcenia chwili na weryfikację podejrzanych lub emocjonalnie nacechowanych treści przed zareagowaniem na nie lub udostępnieniem ich dalej. Oznacza to również sięganie po różnorodne, wiarygodne źródła i gotowość do zmiany poglądów w świetle rzetelnych dowodów.

W istocie bezpieczeństwo poznawcze polega na rozwijaniu praktyk higieny umysłowej w erze deepfake'ów. Podobnie jak uczymy się myć ręce, aby zapobiec rozprzestrzenianiu się zarazków, potrzebujemy zestawu nawyków, praktyk i narzędzi do bezpiecznego poruszania się w środowisku cyfrowym i do takiej konsumpcji informacji, która pozwoli zapobiegać rozprzestrzenianiu się fałszu i manipulacji.

To nie tylko kwestia osobistego dobrostanu, ale także bezpieczeństwa narodowego i globalnego. Technologia obecnie umożliwia złoczyńcom wykorzystywanie informacji jako broni sięjącej niezgodę, zamieszanie i wątpliwości na masową skalę. Dlatego nasza najlepsza obrona będzie polegać na celowym budowaniu odporności poznawczej.

Rzecz w tym, że Twój umysł jest Twoim najcenniejszym zasobem. To soczewka, przez którą postrzegasz rzeczywistość, silnik Twojego procesu decyzyjnego, istota Twojej autonomii. Jego ochrona nie jest tylko aktem troski o siebie, ale aktem samochowawczym w czasach, gdy sama prawda jest zagrożona.

W związku z tym bezpieczeństwo poznawcze to nie tylko miły dodatek, ale niezbędny zestaw umiejętności do poruszania się w świecie cyfrowym z zachowaniem autonomii i integralności. Rozumiejąc słabości naszego poznania i świadomie kształtując nawyki krytycznej konsumpcji i jasnego rozumowania, możemy

wzmocnić nasze umysły przeciwko tym, którzy chcą nas oszukać i zmanipulować.

Ostatecznie celem bezpieczeństwa poznawczego jest wzmocnienie jednostek i społeczeństwa jako całości, abyśmy mogli bronić się przed złośliwymi wpływami, podejmować świadome decyzje i wchodzić w interakcje z innymi ludźmi i światem, opierając się na sprawdzalnych faktach. Jest to kluczowa forma dbania o siebie i odpowiedzialności społecznej w XXI wieku. A wszystko zaczyna się od świadomości.

Podsumowanie

Oszustwa rozprzestrzeniają się z prędkością, a jaką pozwala rozwój technologii. Dlatego ważniejsze niż kiedykolwiek jest zrozumienie zmieniającego się krajobrazu manipulacji cyfrowych i rozwijanie umiejętności poznawczych, które pozwalają się w nim poruszać. Oto trzy kluczowe wnioski:

- **Media syntetyczne**, takie jak deepfaki i treści generowane przez sztuczną inteligencję, stają się coraz bardziej wyrafinowane i trudniejsze do wykrycia. Od fałszywych filmów z celebrytami po sfabrykowane posty polityczne, technologia ta może tworzyć rzeczywistość praktycznie nie do odróżnienia od prawdy. W miarę jak te narzędzia stają się coraz bardziej dostępne i zaawansowane, musimy być przygotowani na przyszłość, w której to, co widzimy, nie zawsze będzie prawdą.
- **Nasze mózgi są zaprogramowane, by ufać temu, co widzimy i słyszymy**, a to czyni nas podatnymi na oszustwa. Złoczyńcy doskonale o tym wiedzą. W pełni to wykorzystują i bombardują nas treściami zaprojektowanymi tak, by wywoływać silne emocje i rozprzestrzeniać się błyskawicznie. Aby zachować trzeźwy osąd, musimy aktywnie doskonalić nasze „detektory

bzdur” oraz rozwijać krytyczne myślenie, świadomość emocjonalną i umiejętność korzystania z mediów.

- **Bezpieczeństwo poznawcze**, czyli ochrona naszych umysłów przed szkodliwymi wpływami i manipulacją, to kluczowa umiejętność w cyfrowym świecie. Przez zrozumienie słabych punktów naszego systemu poznawczego i świadome budowanie odporności psychicznej możemy uzbroić się przeciwko oszustwom i podejmować bardziej świadome decyzje.

Jak więc żyć w świecie pełnym deepfake’ów, dezinformacji i oszustw generowanych przez AI? Należy zacząć od uświadomienia sobie problemu. Musimy zdać sobie sprawę z jego skali i złożoności. Możemy zacząć podejmować kroki, by chronić nasz najcenniejszy zasób: zdolność do jasnego i niezależnego myślenia, nawet gdy jesteśmy zalewani sfabrykowanymi „prawdami” i ukrytymi interesami. Walka o rzeczywistość zaczyna się w umyśle i jest to bitwa, na której przegranie nie możemy sobie pozwolić.

Właśnie dlatego powstała ta książka. Ten pierwszy rozdział to zaledwie muśnięcie tematu. W kolejnych rozdziałach zagłębimy się w działanie generatywnej sztucznej inteligencji, omówimy narzędzia i techniki używane do tworzenia mediów syntetycznych oraz przeanalizujemy szersze implikacje stosowania tych technologii w kreowaniu wyrafinowanych oszustw. Postaramy się też wyjaśnić sposób myślenia atakujących, niektóre z podstępnych metod omijania zabezpieczeń większości systemów AI oraz jak to wszystko wpływa na zaufanie w zakresie osobistym i społecznym.

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 

PRAKTYCZNY PRZEWODNIK PO ŚWIECIE DEEFAKE'ÓW, DEZINFORMACJI I SZTUCZNIE WYGENEROWANYCH OSZUSTW

Przestępcy szybko zrozumieli, jak wsłaniałym narzędziem jest sztuczna inteligencja. Potrafi klonować głos konkretnej osoby, tworzyć fałszywe nagrania wideo i generować przekonujące teksty. Do naszej codzienności wkroczyły oszustwa nowej generacji i wszechobecna dezinformacja. Jak sobie poradzić w świecie ze sztuczną inteligencją i z nieznanymi wcześniej zagrożeniami?

Jeśli chcesz zrozumieć nowe zagrożenia, musisz przeczytać tę książkę.

TED HARRINGTON, lider w świecie etycznego hackingu i cyberbezpieczeństwa

Dzięki tej książce dowiesz się, jak bezpiecznie korzystać z potencjału AI i równocześnie chronić się przed nadużyciami. Na kolejnych stronach odkryjesz niezmiennie prawdy o funkcjonowaniu ludzkiej psychiki i sposoby, w jakie oszuci nią manipulują, korzystając ze sztucznej inteligencji. Uzbroisz się w wiedzę i strategię, które pozwolą Ci prosperować w tej nowej rzeczywistości pełnej nowych niebezpieczeństw. W książce znajdziesz przystępnie przedstawione naukowe wyjaśnienia skuteczności cyfrowych oszustw, a także fascynujące studia przypadków. Poznasz skuteczne strategie ochrony siebie i swoich bliskich przed coraz bardziej wyrafinowanymi metodami przestępców i wykształcisz w sobie świadomość osoby odpornej na manipulację i fałsz.

**Oto pozycja obowiązkowa dla każdego, kto chce zrozumieć,
jak sztuczna inteligencja wpływa na nasze życie.**

FC (FREAKYCLOWN), etyczny haker, autor książki *Jak rabuję banki* (i inne podobne miejsca)

Wzmocnij się i zadbaj o własne bezpieczeństwo!

LISA PLAGGEMIER, National Cybersecurity Alliance

Perry Carpenter jest wielokrotnie nagradzanym autorem, podcasterem i mówcą. Od ponad dwudziestu lat zajmuje się analizą sposobów, w jakie cyberprzestępcy wykorzystują ludzkie zachowania. Jako główny strateg zarządzania ludzkim czynnikiem ryzyka w firmie KnowBe4 tworzy skuteczne systemy obrony przed zagrożeniami opartymi na inżynierii społecznej.

	KOD KORZYŚCI <i>Sięgnij po więcej!</i>  
 helion.pl	ISBN 978-83-289-3107-7
 HELION S.A. ul. Kościuszki 1c 44-100 Gliwice tel.: 32 230 98 63 helion@helion.pl	 9 788328 931077
Cena: 59,90 zł	

WILEY