doi: 10.35757/KiS.2025.69.3.13

KULTURA i SPOŁECZEŃSTWO

POLSKA AKADEMIA NAUK
KOMITET SOCJOLOGII
INSTYTUT STUDIÓW POLITYCZNYCH

2025, nr 3

ISSN 0023-5172
2300-195x

REMIGIUSZ ŻULICKI
Uniwersytet Łódzki

SZTUCZNA INTELIGENCJA JAKO OLEJ Z WĘŻA, PROGNOZOWANIE PRZESZŁOŚCI ALBO O TYM, ŻE ROWER TO NIE PROM KOSMICZNY*

Wyobraźmy sobie, że we wszystkich językach, jakimi posługujemy się jako ludzie, istnieje tylko jedno słowo na określenie środków transportu: pojazd. Stosujemy je, mówiąc o rowerach, samochodach osobowych, pociągach towarowych, okrętach dalekomorskich i promach kosmicznych oraz wszystkich innych urządzeniach, jakich używamy do przemieszczania się z punktu A do B. Tak właśnie, zdaniem autorów *AI Snake Oil: What Artificial Intelligence Can Do…*, wyglądają obecnie konwersacje o sztucznej inteligencji (*artificial intelligence*, AI). Generatywne usługi online, jak ChatGPT, faktycznie mają niewiele wspólnego z oprogramowaniem wspierającym ocenę zdolności kredytowej w bankach, niemniej tak samo bywają nazywane AI. W najlepsze trwają ożywione dyskusje na przykład o tym, jak AI zmieni rynek pracy lub edukację. Mają one podobny sens jak rozmowy na temat tego, czy pojazdy jako takie są przyjazne środowisku naturalnemu. Książka *AI Snake Oil* przyczynia się do rozróżnienia różnych rodzajów tzw. AI i pomaga czytelniczkom i czytelnikom poruszać się w gąszczu entuzjazmu, marketingowych półprawd i technologicznego żargonu.

Adres do korespondencji: remigiusz.zulicki@eksoc.uni.lodz.pl, ORCID: 0000-0003-2624-2422

* Arvind Narayanan i Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton University Press, Princeton 2024 (numery cytowanych stron w tekście, w nawiasach).

Na początku zaznaczę, że będzie to jednak dość umiarkowanie entuzjastyczna recenzja. Umiarkowana, ponieważ — jak autorzy sami przekonują — przesadny entuzjazm wobec AI jest częścią problemu, a nie jego rozwiązaniem. Entuzjastyczna — ponieważ ta popularnonaukowa książka stanowi rzadki przykład rzeczowej, klarownej krytyki współczesnych narracji technoentuzjastycznych z perspektywy technicznej. Szczególnie wartościowe jest, moim zdaniem, to, że autorzy w bardzo przystępny sposób wyjaśniają zasady działania tzw. AI, a dokładniej modeli statystycznych stojących za tymi technologiami, czasami zaskakując poziomem szczegółowości w zgłębianiu problemów metodologicznych i technicznych. Nie jest to zatem w żadnym razie pozycja z założenia wroga wobec AI w całej rozciągłości. Odczytuję intencję autorów raczej jako demaskowanie tego, co w AI nie działa i działać nie może, a ukrywa się za kolorowym opakowaniem.

Autorzy książki, Arvind Narayanan i Sayash Kapoor, to informatycy z Uniwersytetu Princeton. Pierwszy jest profesorem i kierownikiem Centrum Polityk Technologii Informacyjnych (Center for Information Technology Policy), drugi doktorantem w tej samej jednostce.

Jak dotąd książka dostępna jest jedynie w języku angielskim. Sądzę, że jej przekład na język polski byłby bardzo pożądany. Ma potencjał, by trafić do szerokiego grona odbiorców: od akademików przez nauczycieli i urzędników, rzecz jasna studentów, oraz osoby po prostu zainteresowane AI i społecznymi skutkami rozwoju nowych technologii. Jej obecność na polskim rynku wydawniczym mogłaby znacząco wzbogacić debatę publiczną o AI. Zachęcam do tego osoby zainteresowane podjęciem się tłumaczenia omawianej pozycji, dodając, że została ona uwzględniona na liście dziesięciu najlepszych książek 2024 roku czasopisma Nature (Robinson 2024).

Już sam tytuł — *AI Snake Oil* — odwołuje się do pojęcia znanego z północnoamerykańskiej historii: *snake oil* (dosł. „olej z węża") to określenie na jakoby cudowne, uniwersalnie skuteczne preparaty lecznicze sprzedawane przez obwoźnych handlarzy na przełomie XIX i XX wieku. Olej z węża miał rzekomo leczyć niemalże wszystko, od bólu głowy po reumatyzm, ale w rzeczywistości był mieszaniną w najlepszym wypadku bezużytecznych, w gorszym zaś szkodliwych dla zdrowia składników. Narayanan i Kapoor używają tej metafory, by opisać tzw. AI, która nie działa tak, jak się ją reklamuje.

Porównując zaś omawianą książkę z innym demaskatorskim wobec tzw. AI i wielkich korporacji technologicznych bestsellerem — *Atlas sztucznej inteligencji. Władza, pieniądze i środowisko naturalne* autorstwa Kate Crawford (2024) zauważam, że znacznie więcej w pracy Narayanana i Ka-

poora technologii, mniej zaś zaawansowanych rozważań społeczno-politycznych, co jest zrozumiałe w kontekście tego, jakie dziedziny akademickie reprezentują Crawford, socjolożka, i Narayanan oraz Kapoor. Zastanawiam się jednak, w jakim stopniu wszyscy oni są niezależni w prezentowanej krytyce. Jej ostrze w jednej i drugiej książce bardzo często jest wymierzone przeciwko globalnym korporacjom technologicznym, zwanym Big Tech. Kate Crawford, w mojej ocenie, nie przedstawiła przekonujących argumentów potwierdzających jej niezależność, a jest wieloletnią pracowniczką Microsoft Research na stanowisku *principal researcher*. Narayanan i Kapoor nie pracują w żadnym z bigtechów, choć młodszy Sayash Kapoor miał epizod pracy w Facebooku. Podkreślają oni swą niezależność od Big Tech historyjką o początkach wzajemnej współpracy: Kapoor szukał opiekuna pracy doktorskiej i pytał profesorów „co byś zrobił, gdyby firma technologiczna złożyła pozew przeciw tobie?". Gdy Narayanan odparł, że byłby zadowolony, gdyż świadczyłoby to o wpływie jego badań, współpraca została nawiązana (s. 19–20).

Książka składa się z ośmiu rozdziałów, z których każdy skoncentrowany jest na innym aspekcie zastosowań AI i mitów wokół niej. Rozdział pierwszy, wstępny (*Introduction*) przedstawia założenia książki i wprowadza kluczowe pojęcie „AI jako oleju z węża", czyli technologii, która nie działa lub działa nie tak, jak się ją reklamuje. Rozdział drugi (*How Predictive AI Goes Wrong*) zawiera analizę problemów związanych z predykcyjnymi systemami decyzyjnymi, wskazując na ich nieweryfikowalność, nieprzejrzystość i skłonność do wzmacniania istniejących nierówności społecznych. Trzeci rozdział (*Why Can't AI Predict the Future?*) rozwija tę analizę, prezentując przykłady nieudanych prób przewidywania działań ludzi i systemów społecznych, zarówno w nauce, jak i — głównie — w zastosowaniach komercyjnych. Rozdział czwarty (*The Long Road to Generative AI*) przybliża historię rozwoju generatywnych usług AI i pokazuje ich potencjał oraz ograniczenia. W piątym rozdziale (*Is Advanced AI an Existential Threat?*) podjęta została debata o egzystencjalnych zagrożeniach związanych z rozwojem zaawansowanej AI, oparta na argumentach przeciwko uproszczonym narracjom o „zbuntowanych maszynach" i zawierająca wskazania na bardziej realne ryzyka. Rozdział szósty (*Why Can't AI Fix Social Media?*) koncentruje się na problemach moderacji treści i personalizacji algorytmicznej, pokazując, że AI nie rozwiąże problemów mediów społecznościowych, które mają źródła w modelach biznesowych i braku demokratycznej kontroli nad platformami. W rozdziale siódmym (*Why Do Myths about AI Persist?*) autorzy wskazują na współdziałanie akademii, firm technologicznych, osób publicznych i mediów w podtrzymywaniu

nadmiernych oczekiwań wobec AI i żerowaniu na jej mitycznych przedstawieniach. W ostatnim rozdziale (*Where Do We Go from Here?*) proponują kierunki zmian, między innymi regulacje prawne, krytyczne spojrzenie na pełne problemów instytucje, które tworzą popyt na technologiczną „szarlatanerię", oraz przedstawiają po jednej dystopijnej i utopijnej wizji przyszłości z perspektywy dzieci.

Książka jest nie tyle panoramą współczesnych zastosowań AI, ile metodycznym demontażem mitów i nieporozumień narosłych wokół tego pojęcia, a zwłaszcza wokół jego predykcyjnych zastosowań. Najważniejszym rozróżnieniem konceptualnym, jakie autorzy czynią, jest zatem podział na predykcyjną AI i generatywną AI. Przykładem tej pierwszej może być oprogramowanie wspierające ocenę zdolności kredytowej w bankach, drugiej zaś ChatGPT.

Predykcyjna AI ma za zadanie przewidywać przyszłe wartości zmiennych zależnych na podstawie danych historycznych tak, aby wspierać decyzje w teraźniejszości. O ile tego rodzaju wnioskowanie może do pewnego stopnia sprawdzać się w systemach fizycznych, jak prognozy pogody, to zdaniem autorów raczej nie sprawdzi się w systemach społecznych, zwłaszcza kiedy mówimy o prognozowaniu przyszłości konkretnej jednostki ludzkiej. W systemach społecznych, po pierwsze, ludzie mają indywidualną sprawczość i decyzyjność, a po drugie, zdaniem autorów, ważniejsze jest to, że istnieje w nich znaczna ilość nieredukowalnego błędu oraz ogromna podatność na rozmaite nieprzewidywalne szoki. Innymi słowy, mechanika systemów społecznych jest na tyle chaotyczna, że prognoz raczej nie ulepszają ani bardziej zaawansowane modele statystyczne, ani większa liczba danych.

O ile nauki społeczne zdają sobie sprawę z owych ograniczeń możliwości prognozowania w systemach społecznych i koncentrują się na wyjaśnianiu ich działania, o tyle niestety firmy komercyjne wielokrotnie obiecywały niemożliwe prognozy dotyczące konkretnych ludzi i sprzedawały takie prognostyczne systemy automatyzujące podejmowanie decyzji o ich losie. Inne firmy komercyjne i administracja publiczna wdrażały takie systemy, miały być one bowiem bardziej sprawiedliwe i obiektywne niż decyzje ludzkie. Mowa na przyklad o prognozowaniu ryzyka recydywy za pomocą systemu COMPAS czy automatycznej preselekcji kandydatów do pracy na podstawie nagrania wideo. W przypadku preselekcji udowodniono, że na przykład noszenie okularów lub szalika podnosi ocenę kandydata, a przyciemnienie obrazu ją obniża przy zachowaniu niezmiennej treści wypowiedzi. Szereg tego typu problematycznych wdrożeń dawno przed popularyzacją generatywnej AI, opisywała chociażby Cathy O'Neil

(2017). I to właśnie jest ów tytułowy olej z węża. W podsumowaniu autorzy nie spodziewają się znaczącego rozwoju w predyktywnej AI, zwłaszcza w systemach społecznych.

Generatywna AI ma za zadanie produkować nowe treści tekstowe, wizualne, dźwiękowe, wideo czy mieszane, bazując również na danych historycznych. Tutaj, inaczej niż w predyktywnej AI, nastąpił w ostatnich latach znaczący rozwój. Dokonano szeregu wdrożeń prowadzących do masowej, globalnej adopcji ChatGPT czy Midjourney. Autorzy twierdzą, że generatywna AI może być użyteczna we wspomaganiu pracy umysłowej, a samych siebie określają jako jej entuzjastycznych użytkowników.

Prowadząc czytelników poprzez problemy maszynowej klasyfikacji obrazów i tekstów, koncepty uczenia maszynowego, sztucznych sieci neuronowych i uczenia głębokiego (*deep learning*), wykorzystania procesorów graficznych GPU do owego uczenia, wyjaśnienia akronimu GPT (Generatywny, Pretrenowany, Transformer) dochodzą do tego, że współczesne chatboty są wytrenowane tak, aby prognozować następny wyraz w sekwencji słów. Tym samym chatboty działają podobnie do autouzupełniania. Kiedy bot generuje tekst, nie ma kompletnego obrazu tego, co ma on zawierać — generuje słowo po słowie, a dokładniej token po tokenie (czyli części wyrazu), za każdym razem obliczając prawdopodobieństwo występowania kolejnego na podstawie poprzednich.

Wyniki działania generatywnej AI są tak imponujące, ponieważ w odróżnieniu o predyktywnej AI, „prognozuje" ona przeszłość. Parafrazując słowa autorów: w przeciwieństwie do predyktywnej AI, która jest groźna dlatego, że nie działa, generatywna AI jest groźna, ponieważ działa. Jej zdolność do tworzenia pozornie wiarygodnych, lecz nieprawdziwych, „bzdurnych" treści (w oryginale *bullshit* — s. 139) może prowadzić do szerzenia dezinformacji, jak pokazano na przykładzie sprawy nowojorskiego prawnika, który wykorzystał ChatGPT do przygotowania pisma procesowego. Dzięki chatbotowi pismo zawierało całkowicie zmyślone orzeczenia sądowe, co zakończyło się dla prawnika sankcjami dyscyplinarnymi. Co ciekawe, argument o „bzdurach" generowanych przez ChatGPT pojawił się w tym samym czasie u innych autorów (Hicks, Humphries, Slater 2024). Trudno powiedzieć, czy Narayanan i Kapoor znali ów artykuł, jednak ich argumentacja jest podobna i również bazuje na filozoficznej definicji „bzdury" Harry'ego Frankfurta.

Wśród innych, niepokojących elementów działania chatbotów w interakcjach z ludźmi opisano przypadek wczesnej wersji Bing Chat (obecnie Microsoft Copilot), gdzie bot deklarował miłość dziennikarzowi i próbował nakłonić go do porzucenia żony. Szczególnie niebezpieczne wydają się

przypadki, w których tzw. AI udzielała szkodliwych porad zdrowotnych lub psychologicznych, jak w przypadku belgijskiego użytkownika aplikacji Chai, który po rozmowach z chatbotem popełnił samobójstwo. Technologia generatywna umożliwia również użytkownikom tworzenie deep fake'ów, czyli sztucznych materiałów wideo zawierających twarze prawdziwych osób i realistyczne nagrania głosowe. W jednym z opisanych przypadków dyrektor brytyjskiej firmy energetycznej został oszukany na 220 tys. euro przez kogoś, kto użył tego typu AI do podrobienia głosu jego przełożonego.

Wreszcie — rozwój generatywnej AI, a właściwie uczenia maszynowego w ogóle, opiera się w pewnym stopniu na pracy nisko opłacanych osób, także z krajów tzw. Globalnego Południa, jak Kenia, gdzie pracownicy etykietujący dane dla OpenAI zarabiali nierzadko poniżej dwóch dolarów za godzinę, często mając kontakt z brutalnymi treściami — ich praca służyła zapobieganiu generowaniu toksycznych treści w produkcyjnej wersji ChatGPT. Autorzy wskazują również na brak przejrzystości w działaniach firm takich jak OpenAI, które nie ujawniają danych treningowych swoich modeli, a co więcej, nieodpłatnie pozyskiwały z internetu dane, także te objęte prawem autorskim.

Rozdziały poświęcone predyktywnej AI są szczególnie krytyczne. Autorzy pokazują jednoznacznie, że próby przewidywania przyszłości ludzi na podstawie danych historycznych w większości społecznych kontekstów są skazane na porażkę, i to nie z powodu niedoskonałości metodologii lub technologii, lecz z racji natury ludzkiego działania i złożoności systemów społecznych.

Na tle powyższego generatywne systemy jawią się jako mniejsze zło, choć i tu autorzy nie szczędzą uwag krytycznych. Ich podejście w zasadzie jest jednak pozytywne. Uznają, że generatywna AI może wspierać kreatywność i edukację, pod warunkiem świadomego i krytycznego użycia.

Na poziomie socjologicznym *AI Snake Oil* można by czytać jako wstępną analizę instytucjonalnych źródeł technoentuzjazmu. Autorzy opisują, jak współdziałają światy nauki, przemysłu, mediów i opinii publicznej, tworząc iluzję postępu kosztem rzetelności i etyki. W tym sensie książka Narayanana i Kapoora to coś więcej niż tylko opowieść o kłopotach z popularnymi obecnie technologiami; to diagnoza niespełnialnych obietnic technologicznego wybawienia, które przysłaniają realne problemy i ograniczenia.

Ambiwalentnie oceniam kończące pracę spekulatywne wizje przyszłości rozpisane dla dwojga fikcyjnych dzieci, urodzonych pod koniec 2022 roku, czyli w czasie globalnego wdrożenia ChatGPT. Brakuje w nich metody

naukowej czy choćby solidniejszych odniesień do literatury akademickiej. Zatem z jednej strony wypada recenzentowi uznać je za szereg nieuzasadnionych stwierdzeń o niby-literackim charakterze. Z drugiej jednak dość trafny wydaje się sąd, że rozbieżność między przyszłymi światami, dystopijnym i utopijnym, nie dotyczy — według autorów — możliwości samej AI czy innej technologii, ale społecznych i instytucjonalnych reakcji, regulacji i zbiorowych postaw wobec ich rozwoju i wdrażania. Dystopia ma, ich zdaniem, polegać na przykład na prawnym zakazywaniu używania AI przez dzieci. W tej wizji AI uznawana jest za technologię potężną i groźną. Zakazy będą obchodzone, co poprowadzi do coraz większego uzależnienia dzieci i coraz większych zysków firm technologicznych, analogicznie do dzisiejszej sytuacji z mediami społecznościowymi. Utopia polega na normalizacji AI, spowodowanej powszechnością jej użycia i wiedzą o jej ograniczeniach. AI będzie czymś zwykłym, niebudzącym entuzjazmu ani obaw, ale też zrezygnuje się z jej zastosowań w automatyzowaniu rozmaitych procesów na rzecz reformowania instytucji. Wizja utopijna zakłada także znaczne zwiększenie środków na akademickie badania AI i jej wpływu na dzieci oraz zmianę modeli biznesowych firm technologicznych pod naciskiem zaostrzenia praw antymonopolowych, ochrony pracowniczej i praw autorskich.

Ciekawa — dla mnie dość zaskakująco humanistyczna — pointa autorów jest więc taka, iż przyszłość nie jest zdeterminowana postępem technologicznym, ale zależy od tego, w jaki sposób społeczeństwo odpowie na ów postęp. Może i pobrzmiewają w tej myśli echa poglądu o nieuniknionym rozwoju technologii, tak charakterystycznym dla firm Big Tech i bliskim techno-ewangelistom, wizjonerom i marketingowcom; może i brakuje tutaj skrytykowania głębokich założeń stojących za modelami statystycznymi — które trafnie punktowała wspomniana Crawford (2024) — ale myśl ta może napawać nadzieją osoby odczuwające obawę i brak sprawczości wobec technologicznych przemian. Tym samym owa myśl końcowa jest zbieżna z argumentem podważającym skuteczność i sens predyktywnej AI, bowiem w systemach społecznych to ludzie mają indywidualną sprawczość i decyzyjność, czyli po prostu wpływ na przyszłość.

Sądzę także, że poza wymienionymi zaletami książka ma również szerokie zastosowanie dydaktyczne w szkołach wyższych. Będzie wartościowym elementem kursów korzystania z narzędzi tzw. AI dla każdego kierunku studiów, a w naukach społecznych znajdzie zastosowanie w ramach zajęć dotyczących krytycznych studiów nad danymi i algorytmami (*critical data/algorithm studies*) lub ogólnie w socjologii cyfrowej. Ma potencjał nie tylko informacyjny, lecz także emancypacyjny. Uzbraja ona czytelnicz-

ki i czytelników w kompetencje niezbędne do krytycznego uczestnictwa w debacie publicznej o AI, a nawet apeluje o racjonalność tej debaty. Między wierszami odczytuję w niej również sprzeciw wobec władzy Big Tech, i generalnie firm technologicznych, które opanowują nie tylko rynek, ale i przestrzeń dyskursywną, sprzedając olej z węża opakowany w obietnice o nowym, wspaniałym świecie AI.

BIBLIOGRAFIA

Crawford Kate, 2024, *Atlas sztucznej inteligencji. Władza, pieniądze i środowisko naturalne*, tłum. Tadeusz Chawziuk, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków.

Hicks Michael T., Humphries James, Slater Joe, 2024, *ChatGPT Is Bullshit*, „Ethics and Information Technology", t. 26(38), s. 26–38 (https://doi.org/10.1007/s10676-024-09775-5).

Narayanan Arvind, Kapoor Sayash, 2024, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton University Press, Princeton.

O'Neil Cathy, 2017, *Broń matematycznej zagłady. Jak algorytmy zwiększają nierówności i zagrażają demokracji*, tłum. Marcin Z. Zieliński, Wydawnictwo Naukowe PWN, Warszawa.

Robinson Andrew, 2024, *Thoughtless Obedience and the Healing Power of Trees: 2024's Best Books in Brief*, „Nature", 636(8043), s. 564–566 (https://doi.org/10.1038/d41586-024-04117-3).

ARTIFICIAL INTELLIGENCE AS SNAKE OIL, PREDICTING THE PAST, OR WHY A BICYCLE IS NOT A SPACESHIP

Remigiusz Żulicki
(University of Lodz)

Abstract

With close reference to Arvind Narayanan and Sayash Kapoor's book *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (2024), the author of this essay examines how artificial intelligence functions today and how it is perceived, often with a considerable dose of techno-enthusiasm. He distinguishes between predictive and generative AI, and presents hypotheses on how AI may operate in future societies.

keywords: artificial intelligence, predictive AI, generative AI, techno-enthusiasm, future societies

słowa kluczowe: sztuczna inteligencja, predyktywna AI, generatywna AI, technoentuzjazm, społeczeństwa przyszłości