Sztuczne inteligencje i biologiczne mózgi*
Artifical Inteligences and Biological Brains

Piotr Durka**

Wydział Fizyki UW

Abstrakt. Sztuczna inteligencja (ang. *artificial intelligence*, AI) to najgorętszy temat ostatnich lat, nie tylko w technologii. Jest wszędzie – od szczoteczek do zębów po artykuły naukowe. Pochłania setki miliardów dolarów, trzęsie giełdami, podważa wiarę w prawdziwość cyfrowych treści, halucynuje i karmi apokaliptyczne przepowiednie. Czym naprawdę jest AI? Czy zamiast *Artificial Intelligence* powinniśmy mówić o *Alien Intelligence*, jak sugeruje Yuval Noah Harari, czy raczej oczekiwać połączenia inteligencji białkowej z krzemową przez interfejsy mózg-komputer, razem z Raymondem Kurzweilem? Dlaczego wykorzystująca zdobycze nauki cywilizacja skręca nagle w stronę czarnych skrzynek i tajemniczych wyroczni? Spróbujemy określić, czym jest AI, i wyjaśnimy czym nie jest, demaskując po drodze kilka miejskich legend o podsłuchiwaniu myśli i przenoszeniu świadomości do cyberprzestrzeni. Omówimy też realne zagrożenia wynikające z faktu, że od od lat oddajemy algorytmom rząd dusz, ale nie zauważamy tego wsłuchani w opowieści o *nadchodzącej* „apokalipsie AI".
Słowa kluczowe: sztuczna inteligencja, sztuczna sieć neuronowa, interfejs mózg-komputer, uczenie maszynowe, algorytm, media społecznościowe

Abstract. Artificial Intelligence (AI) is the hottest topic of recent years, and not only in technology. It is everywhere – from toothbrushes to scientific articles. It consumes hundreds of billions of dollars, shakes stock markets, undermines the credibility of digital content, hallucinates and feeds apocalyptic prophecies. What is AI really? Should we understand Artificial Intelligence as *Alien Intelligence*, as Yuval Noah Harari suggests, or rather expect biological intelligence to merge with silicon intelligence via brain-computer interfaces, together with Raymond Kurzweil? Why does our science-based civilization suddenly turn towards black boxes and mysterious oracles? We will try to define what AI is and explain what it is not, along the way debunking a few urban legends about eavesdropping on thoughts and transferring consciousness to cyberspace. We will also discuss the real threats resulting from the fact that for years we have been giving the reign of our souls to algorithms, but we do not notice it, listening to stories about the *coming* "AI apocalypse".
Key words: artificial intelligence, artificial neural network, brain-computer interface, machine learning, algorithm, social media

1. Czy to jest AI?

Według wdrażanego od 2 lutego 2025 Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024/1689 [21]:

„system AI" oznacza system maszynowy, który został zaprojektowany do działania z różnym poziomem autonomii po jego wdrożeniu oraz który może wykazywać zdolność adaptacji po jego wdrożeniu, a także który – na potrzeby wyraźnych lub dorozumianych celów – wnioskuje, jak generować na podstawie otrzymanych danych wejściowych wyniki, takie jak predykcje, treści, zalecenia lub decyzje, które mogą wpływać na środowisko fizyczne lub wirtualne [...]

Przyjrzyjmy się algorytmom i metodom obliczeniowym, które wydają się pasować do tej definicji.

1.1. Statystyka

Przykładem wspomnianych w powyższej definicji *decyzji* może być ocena zdolności kredytowej. Zadaniem szerokiej klasy metod statystycznych jest określenie, czy nowego klienta zaliczyć do grupy spłacających kredyty (zielone punkty na rys. 1), czy niespłacających (czerwone punkty). Granicę między tymi dwoma grupami wyliczamy na podstawie historii kredytowych poprzednich klientów banku, określanych w języku uczenia maszynowego zbiorem uczącym. Procedura jej wyznaczenia za pomocą znanej od roku 1936 liniowej analizy dyskryminacyjnej (ang. *linear discriminant analysis*, LDA [7]) jest stosunkowo prosta i jednoznaczna. Jakość przewidywania zależy wyłącznie od ilości i jakości danych wejściowych służących estymacji granicy między grupami. Po jej wyznaczeniu (czerwona linia na rys. 1) możemy już szybko i łatwo ocenić, do której z grup prawdopodobnie będzie należał nowy klient.

*Wersję artykułu do publikacji Redakcja PF otrzymała 24 lutego 2025.
**ORCID 0000-0001-5816-8082

Rys. 1. Zielone i czerwone punkty oznaczają klientów, którzy w przeszłości odpowiednio spłacali kredyty lub nie. Czerwona linia to podział, według którego potencjalny kredytobiorca będzie przypisany do jednej z grup

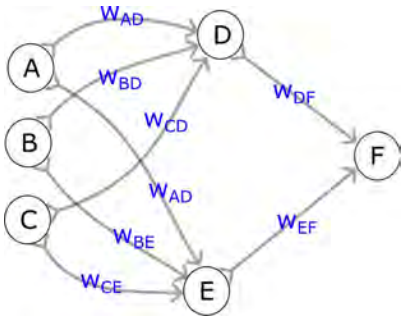
Prosty system, który co jakiś czas będzie dodawał dane o spłacalności ostatnio udzielanych kredytów do zbioru uczącego i wyliczał od nowa graniczną linię za pomocą tego samego wzoru, spełnia warunek adaptacji po wdrożeniu, generuje decyzje, a jego autonomia zależy wyłącznie od woli wdrażających. Czy to już jest AI?

1.2. Sztuczne sieci neuronowe

Jeśli nie chcemy się ograniczać do podziałów liniowych, możemy skorzystać z nieliniowych analogów LDA lub na przykład sztucznych sieci neuronowych (ang. *artificial neural networks*, ANN), które nieliniowość mają wbudowaną we wszystkie przetwarzające informację węzły. Każdy węzeł na wyjściu zwraca funkcję ważonej sumy wejść; na przykład wyjście węzła D na rys. 2 wyniesie

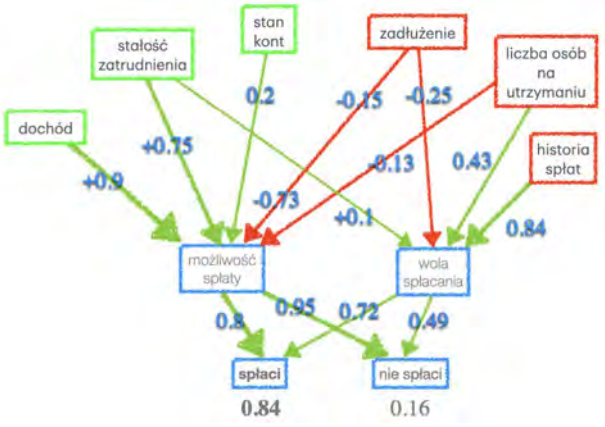
$$D = f(w_{AD}A + w_{BD}B + w_{CD}C), \qquad (1)$$

gdzie f to funkcja nieliniowa (np. sigmoida).



Rys. 2. Obliczenia w sztucznej sieci neuronowej: po podaniu wartości wejścia w węzłach A, B i C, obliczane są według wzoru (1) wartości w węzłach D i E, a na koniec wartość wyjściowa F

Sieć oceniająca zdolność kredytową na podstawie sześciu parametrów mogłaby wyglądać tak, jak przykład z rys. 3. Klasyfikacja nowego przypadku ogranicza się do kilkukrotnego zastosowania wzoru (1), a wynik odczytujemy z wartości węzłów ostatniej warstwy. Obliczenia te



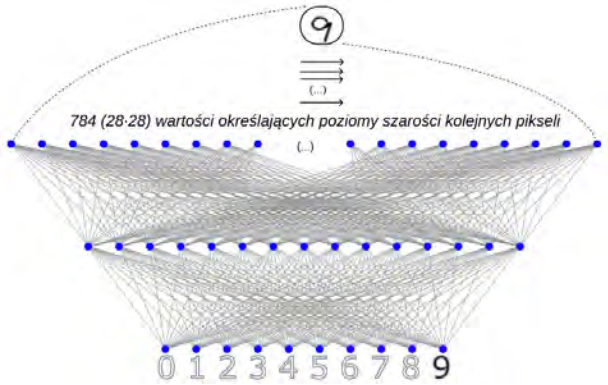
Rys. 3. Hipotetyczna sieć oceniająca zdolność kredytową

są bardzo proste, kiedy znamy wagi połączeń w_{XY}. Dopasowanie tych wag do zbioru uczącego jest już bardziej skomplikowane – algorytm propagacji wstecznej znamy od niespełna pół wieku. Schemat zastosowania nie odbiega od przykładu z poprzedniego rozdziału: co jakiś czas, gdy napłyną nowe dane, uczymy sieć od początku i pozwalamy generować decyzje. Czy to już jest AI?

1.3. Uczenie głębokie

Sieć na rys. 3 jest względnie prosta: znając wagi w_{XY}, obliczenia potrzebne do klasyfikacji nowego wejścia możemy wykonać na kartce, a dwóm węzłom w środkowej („ukrytej") warstwie możemy nawet próbować przypisywać znaczenia odpowiadające stadiom pośrednim procesu klasyfikacji.

Spróbujmy zmierzyć się z większymi wymiarami wejścia, na przykładzie rozpoznawania pisanych odręcznie cyfr (rys. 4). Warstwa wejściowa musi mieć wymiar odpowiadający rozmiarom analizowanych obrazów, a na wyjściu chcemy otrzymać rozróżnienie na 10 klas (cyfry od 0 do 9).



Rys. 4. Schemat sieci rozpoznającej odręcznie pisane cyfry

Funkcja klasyfikująca obrazy jest zapisana w wagach w_{XY}, więc dla nietrywialnej klasyfikacji liczba tych parametrów nie może być zbyt mała. Liczba parametrów sieci stosowanych do rozpoznawania cyfr z bazy MNIST (zawierającej 60 tysięcy przykładów odręcznie pisanych

cyfr) rosła na przełomie wieków od kilkuset do nawet miliona, dając błędy rozpoznawania na poziomie ułamków procenta.

Milion parametrów to dzisiaj bardzo mało – do klasyfikacji bardziej złożonych obrazów sieci neuronowe potrzebują znacznie większej liczby parametrów. Im więcej parametrów, tym więcej zasobów wymaga uczenie i działanie sieci i tym trudniej interpretować ich znaczenie.

Boom na „naprawdę głębokie" sieci neuronowe [8] zapoczątkowała sieć AlexNet, złożona z 650 tysięcy węzłów i 60 milionów parametrów. W roku 2012 uzyskała wyniki znacznie przewyższające wszystkie dotychczasowe podejścia z zakresu rozpoznawania obrazów (ang. *computer vision*). Przełom spowodowały przede wszystkim dwa, do dzisiaj kluczowe, czynniki:

- Dostępność (w Internecie) ogromnej liczby zdjęć, dzięki czemu prof. Fei Fei Li doprowadziła do powstania ImageNet – zbioru milionów obrazów z oznaczeniami treści, nadającego się do uczenia nadzorowanego [5]. Sieci neuronowe uczone na mniejszej liczbie danych dawały rezultaty gorsze niż klasyczne metody rozpoznawania obrazów.
- Dostępność ogromnych mocy obliczeniowych, w szczególności specjalizowanych procesorów do obliczeń graficznych (ang. *graphical processing units*, GPU), których masowo równoległa architektura przyspieszyła proces uczenia sieci i umożliwiła stosowanie znacząco większych liczb parametrów.

Współczesne sieci neuronowe klasyfikują obrazy nie gorzej od ludzi. Przykładowym zastosowaniem, opisywanym w mediach pod hasłem „AI leczy raka", jest detekcja nowotworów w obrazach radiologicznych. Odpowiednio duża sieć w krótkim czasie nauki (czyli dostosowywania wag w_{XY}) może „przejrzeć" więcej obrazów niż radiolog przez całe życie, uzyskując „nadludzką" (lub nie gorszą od eksperta) dokładność. Statystycznie.

Przy tak ogromnych rozmiarach sieci, określenie, które cechy obrazów są wykorzystywane w klasyfikacji, jest niezmiernie trudne lub niemożliwe; dąży do tego, na razie bez wielkich sukcesów, dziedzina po angielsku zwana *explainable AI*. Jednak nawet traktując sieci jak czarne skrzynki można pokazać, że w procesie klasyfikacji nie wykorzystują one tych samych cech obrazów, co ludzie. Ilustrują to zjawisko tzw. ataki jednego piksela [23]. Okazuje się, że zmiana jednego (!) piksela w obrazie RTG zdrowego płuca może zmienić klasyfikację sieci na „zapalenie płuc" i odwrotnie [24]. Jest to ilustracja paradoksu Moraveca [20]: rzeczy proste dla ludzi bywają niezmiernie trudne do odtworzenia przez komputery i odwrotnie. Ekspert do klasyfikacji obrazów radiologicznych wykorzystuje całe swoje wykształcenie, doświadczenie i rozumienie, czym jest zdjęcie, podczas gdy sieć wykorzystuje tylko statystyczne różnice w grupach pikseli, co w większości przypadków wystarcza do klasyfikacji. Czy to już jest AI?

1.4. Generatywna AI

Największą eksplozję zainteresowania AI spowodowało udostępnienie w Internecie interfejsów umożliwiających konwersację w języku naturalnym z Dużymi Modelami Językowymi (ang. *large language models*, LLM).[1] Dla AI stały się one tym, czym WWW w ostatniej dekadzie ubiegłego wieku było dla Internetu: umożliwienie korzystania z usług internetowych osobom nie posiadającym wiedzy specjalistycznej i nie rozumiejącym zasad ich działania spowodowało gigantyczny wzrost zainteresowania oraz inwestycji. Analogicznie dzisiaj każdy może „porozmawiać z AI" i formułować na tej podstawie własne opinie. Okazuje się, że napisanie przez LLM sensownego eseju, który można przedstawić jako pracę domową, robi zdecydowanie większe wrażenie niż przewidywanie struktury białek przez model AplhaFold [12], za co przyznano nagrodę Nobla z chemii w 2024 roku.

Dzięki czatom z LLM opinie o AI są najczęściej entuzjastyczne, a przeważająca sensowność odpowiedzi uruchamia wrodzoną ludziom skłonność do antropomorfizacji. Na przykład, pojawiające się od czasu do czasu w generowanych przez LLM tekstach, kompletne bzdury, określamy mianem halucynacji lub kłamstw, choć m.in. według autorów artykułu *ChatGPT is bullshit* [9] określenia te nie mają sensu w odniesieniu do bytów „nieznających" pojęcia prawdy, tylko produkujących teksty przypominające stwierdzenia prawdziwe. Są one generowane na podstawie statystycznych własności tekstów pobieranych z Internetu, bez uprzedniej selekcji ani weryfikacji. Same modele również nie mają wbudowanych żadnych mechanizmów sprawdzania prawdziwości i między innymi dlatego bywają przez naukowców nazywane stochastycznymi papugami [2].

„Poziom inteligencji" AI sprawdzamy za pomocą testów i zadań, których w Internecie nie brakuje. I znów, w zdecydowanej większości przypadków, LLM rozwiązują kolejne testy „z nadludzką dokładnością", co staje się pretekstem do medialnych doniesień o tym, że AI przekroczyło właśnie kolejny poziom – ośmiolatka lub doktoranta. Z ekstrapolacji tak postrzeganego trendu na kolejne lata wynikają prognozy o bliskim końcu świata rządzonego przez ludzi itp. Jednak jeśli przyjrzeć się bliżej, wyraźnie widać różnicę między studentem, który tylko przeczytał zbiory zadań z odpowiedziami, a takim,

1. Pierwsze programy komputerowe prowadzące proste konwersacje w języku naturalnym powstawały już ponad pół wieku temu. Najbardziej rozpoznawalnym była ELIZA (nazwa nawiązuje do sztuki Pygmalion), opisana w artykule z 1966 roku [26].

który chociaż próbował zadania rozwiązać. Autorzy artykułu [19] napotykają ślady tej pierwszej sytuacji: wyraźny spadek wyników w obecności drobnych modyfikacji standardowych testów i zadań (np. zmiana występujących w tekstach zadań imion czy liczb) sugeruje, że współczesne[2] LLM w miejsce przypisywanego im logicznego rozumowania odtwarzają, drogą statystycznego dopasowywania wzorców, kroki obecne w danych uczących.

Niezależnie od tego, czy procesy te uznamy za logiczne myślenie, czy nie, nie zachodzą one w tak prostych systemach jak ANN opisywane wcześniej, dlatego nie mówimy już o sieciach, tylko o modelach. Współczesne LLM wykorzystują architekturę transformerów (stąd nazwa ChatGPT, ang. *generative pre-trained transformer*), opisaną po raz pierwszy w roku 2017 w artykule *Attention is all you need* [25], i wiele innych błyskotliwych technik matematycznych, których omówienie wykracza poza ramy tego artykułu. Za intuicyjny przykład posłużyć może technika uczenia polegająca w przybliżeniu na podawaniu sieci na wejściu zdań, z których usunięto (np. ostatnie) słowo, i dopasowywaniu wag tak, aby to właśnie słowo pojawiło się na wyjściu, czyli takie „autouzupełnianie na sterydach" odzwierciedlające statystyczne własności wszystkich tekstów świata. Na wyjściu model dobiera słowa na podstawie prawdopodobieństw, ale z elementem losowym, więc reakcja na dane pytanie nie zawsze będzie jednakowa. To już chyba, wedle aktualnych przekonań, jest AI.

2. Emulacja[3] mózgu

Opisane w rozdziale 1.2 węzły ANN (sztucznych sieci neuronowych) zwiemy zwykle neuronami, co może prowadzić do nadinterpretacji i nieporozumień. Co ANN i wykorzystujące je AI, mają wspólnego z mózgiem?

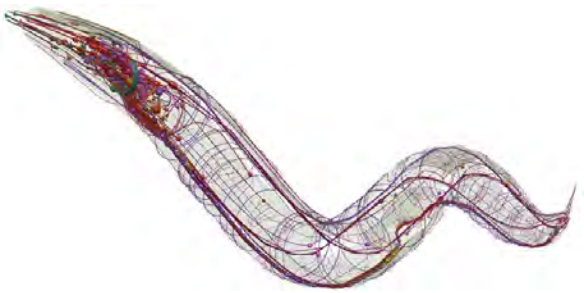
W roku 1943, w słynnym artykule *A logical calculus of ideas immanent in nervous activity* [16], McCulloch i Pitts zaproponowali prosty model neuronu, aby wykazać, że złożone z takich jednostek sieci mogą wykonywać operacje logiczne i obliczenia jak maszyna Turinga. Nie chodziło tu o symulacje działania mózgu, model wykorzystywał bowiem dość luźno ówczesny stan wiedzy o układzie nerwowym, a w referencjach znalazły się tylko trzy (!) prace z zakresu logiki formalnej.

Węzły obliczeniowe współczesnych ANN różnią się tylko nieznacznie od wersji z roku 1943. W modelu McCullocha i Pittsa wszystkie połączenia miały te same wagi, a jedno połączenie hamujące wygaszało całkowicie możliwość wygenerowania potencjału w danym cyklu (veto). Współczesne węzły ANN do sumy wejść pobudzających i hamujących (o dodatnich i ujemnych wagach w_{XY}) stosują funkcję aktywacji f z równania (1), której postać nie jest dobierana pod kątem zgodności z neurobiologią, tylko efektywności obliczeniowej ANN. Postęp wyznaczają rosnące rozmiary sieci.

Całkiem inaczej wygląda postęp w zupełnie odrębnej dziedzinie modelowania czynności neuronów biologicznych. Już w roku 1952 Hodgkin i Huxley zaproponowali układ nieliniowych równań różniczkowych [11], uwzględniających przepływ jonów sodu i potasu przez błonę neuronu, opisujący powstawanie obserwowanych w neuronach potencjałów czynnościowych (za co 10 lat później otrzymali nagrodę Nobla). Parametry modelu były dopasowywane do wyników eksperymentalnych. Kolejne dekady postępów w neurobiologii dały ogromny materiał, pozwalający na tworzenie niemal dowolnie dokładnych modeli matematycznych neuronów biologicznych i symulowanie ich interakcji, na przykład dla zrozumienia podłoża chorób neurologicznych. Do dzisiaj symulujemy w tym celu ich mniejsze lub większe zespoły, dobierając stopień złożoności do pytań badawczych. Ale nie cały mózg.

Próby całościowej symulacji kompletnych organizmów najlepiej oddają trwające od ponad ćwierćwiecza prace nad układem nerwowym nicienia *Caenorhabditis elegans* (*C. elegans*) (rys. 5). Dlaczego akurat ten malutki (ok. 1 mm) robaczek stał się tak popularny w neuronaukach? Badania nad tym organizmem są nieporównanie łatwiejsze niż badania na ludziach:

- konektom (czyli kompletny schemat połączeń neuronów) *C. elegans* znamy od roku 1986 [27],
- układ nerwowy *C. elegans* składa się dokładnie z 302 neuronów – mózg człowieka z ponad 86 miliardów,
- wszystkie robaczki tego gatunku mają taki sam konektom – mózg każdego człowieka jest inny,
- konektom *C. elegans* jest niezmienny – neuroplastyczność mózgu człowieka powoduje, że połączenia między neuronami (i same neurony) powstają i giną.



Rys. 5. Wizualizacja układu nerwowego nicienia *Caenorhabditis elegans*, wygenerowana na stronie http://browser.openworm.org

2. Artykuł [19] opublikowano w październiku 2024, tuż przed pojawianiem się modelu OpenAI o1.

3. Emulacja od łac. *aemulatio* (naśladować); w informatyce technika rozpoznawania przez układ elektroniczny lub program (zwany emulatorem) danych przeznaczonych dla innego układu lub programu (przyp. red.).

Ale droga od konektomu do odtworzenia choćby podstawowych zachowań (nie mówiąc o świadomości) jest co najmniej bardzo długa. Nawet w przypadku tak prostego organizmu jak *C. elegans* jesteśmy wciąż dopiero na jej początku, co pokazują na przykład artykuły podsumowujące dyskusję *Connectome to behaviour: modelling C. elegans at cellular resolution* [22]. Inaczej mówiąc, przeniesienie układu nerwowego malutkiego robaczka do cyberprzestrzeni tak, żeby odtwarzać choćby jego podstawowe zachowania, nie jest aktualnie możliwe i nie można uczciwie powiedzieć, czy i kiedy będzie możliwe. W tym właśnie kontekście należy oceniać powracające w mediach zapowiedzi emulacji *ludzkiego* mózgu i transferu umysłu do cyberprzestrzeni.

3. Błąd ekstrapolacji

Bezpośrednie porównywanie systemów AI do mózgu jest równie sensowne, jak nazywanie samolotów sztucznymi ptakami – nie oczekujemy, że dojrzałe technologie lotnicze dadzą nam samoloty znoszące jajka. Analogicznie, współczesne systemy AI konstruowane są w celu wykonywania konkretnych zadań, a nie w celu poznawania i odtwarzania działania ludzkiego mózgu.

Sztandarowym projektem w tej drugiej dziedzinie był Human Brain Project, który, pomimo finansowania na poziomie miliarda euro, nie spełnił obietnicy Henry'ego Markrama, wyrażonej pod koniec wykładu na konferencji TED[4] w roku 2009 [15]:

[...] *mam nadzieję, że przynajmniej częściowo przekonałem was, że zbudowanie mózgu nie jest niewykonalne. Możemy to zrobić w ciągu 10 lat i jeśli się nam powiedzie, wyślemy do TED, za 10 lat, hologram, który z wami porozmawia.*

Skąd się biorą tak nierealistyczne obietnice? Naiwnie można by je wytłumaczyć błędem ekstrapolacji: Rozumiemy już dość dokładnie działanie pojedynczych neuronów i interakcje między nimi. Rozumiemy, czyli potrafimy zasymulować numerycznie. Działanie mózgu opiera się na interakcjach między grupami neuronów. Potrafimy już symulować wybrane aspekty działania niewielkich grup neuronów, na przykład na potrzeby badań nad epilepsją. Mogłoby się wydawać, że jeśli tylko uruchomimy odpowiednio potężny komputer, który pozwoli na efektywną symulację coraz większych grup neuronów, to w pewnym momencie, z samej skali, automagicznie wyłoni się nowa jakość, czyli świadomość i komputer znienacka ogłosi: *Cześć, jestem Ambroży*. Albo wręcz od razu [18]:

Daj mi rząd dusz!–Tak gardzę tą martwą budową
Którą gmin światem zowie i przywykł ją chwalić,
Żem nieprobował dotąd czyli moje słowo,
Niemogłoby jej wnet zwalić.
Lecz czuję w sobie, że gdybym mą wolę
Ścisnął, natężył i razem wyświecił,
Może bym sto gwiazd zgasił, a drugie sto wzniecił.

Podobnie w dziedzinie AI, skala może się wydawać najważniejszym parametrem. Przecież, jak pisaliśmy w rozdziale 1.3, sztuczne sieci neuronowe rozwinęły skrzydła dopiero dzięki wystarczająco ogromnej skali rozmiaru zbiorów uczących i zasobów obliczeniowych. Dlatego setki miliardów dolarów i budowanie dedykowanych elektrowni atomowych dla zaspokojenia potrzeb centrów obliczeniowych mają automatycznie doprowadzić do powstania skali, w której AI stanie się wszechwiedzącą wyrocznią, najpotężniejszą bronią i źródłem niewyobrażalnego bogactwa. Podobnie jak opowieści Henry'ego Markrama o sztucznym mózgu zapewniły miliard euro na prowadzone przez niego badania, tak historie o nadludzkiej mocy, którą już niedługo osiągnie AI, przynoszą dziś setki miliardów inwestycji w te technologie.[5]

Kruchość tej bańki pokazało niedawne upublicznienie przez chińską firmę DeepSeek nowego modelu [1], który według twórców wymaga znacząco mniejszych zasobów obliczeniowych niż wiodące modele tworzone w USA. Firma NVIDIA, produkująca kluczowe dla uczenia dużych modeli układy scalone, zanotowała największy w historii amerykańskiej giełdy spadek wartości o prawie 600 miliardów dolarów; firma OpenAI (która wbrew nazwie nie ma nic wspólnego z otwartością) oskarża DeepSeek o niezgodne z licencją wykorzystanie ich modelu (ChatGPT) w procesie uczenia chińskiego modelu R1; autorzy i dziennikarze oskarżają OpenAI o kradzież własności intelektualnej z powodu „karmienia" modeli danymi objętymi prawami autorskimi…

4. Jeśli nie możesz jej pokonać, przyłącz się przez BCI?

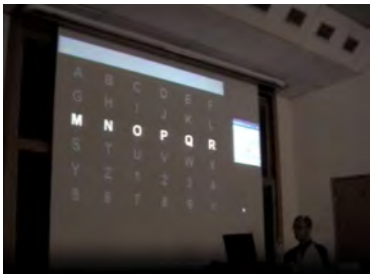
Kolejnym sposobem, w jaki technologiczni celebryci obiecują uratować ludzkość przed apokalipsą AI, jest połączenie ludzkiego mózgu z krzemowym. Z pozoru prosta sprawa: interfejsy mózg-komputer (ang. *brain-computer interfaces*, BCI) istnieją od lat, jacyś naukowcy nad tym pracują, więc pewnie wystarczy im dorzucić parę milionów i załatwione, w czym więc problem? Wyjaśnijmy to dokładniej:

BCI powstały w ubiegłym wieku z myślą o cierpiących na choroby neurodegeneracyjne, takie jak np.

4. TED (ang. *Technology, Entertainment, Design*), to marka konferencji naukowych organizowanych corocznie przez amerykańską organizację *non-profit* Sapling Foundation, celem których jest popularyzacja idei wartych propagowania (przyp. red.).

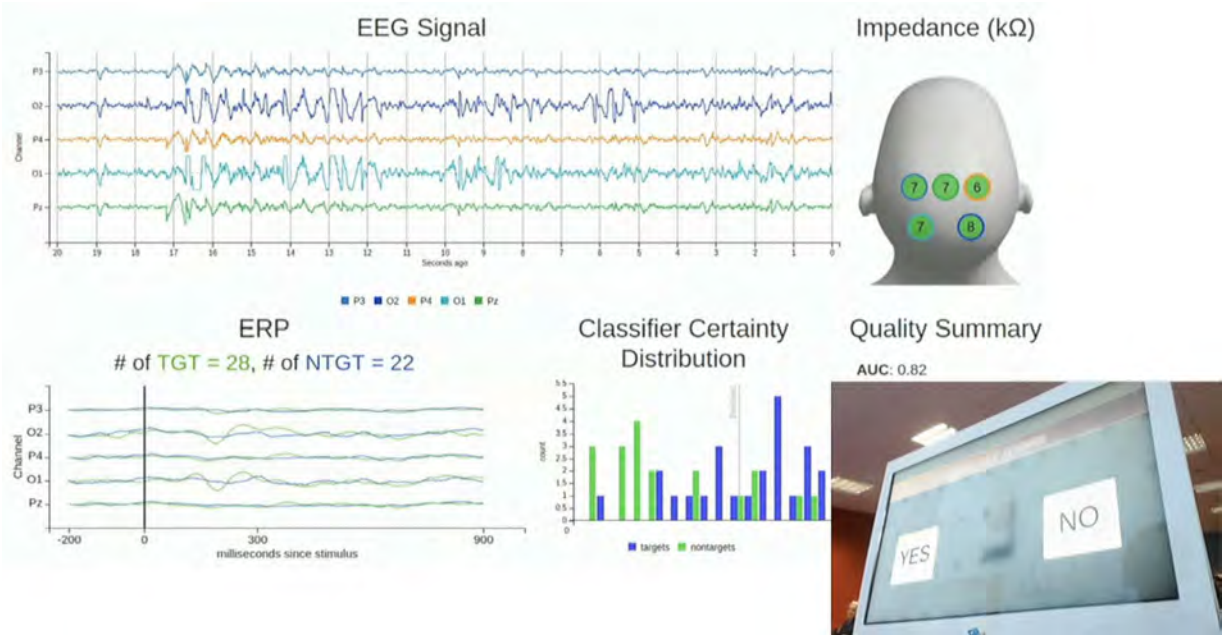
5. Obiektywność opinii i inne kwestie związane z AI dyskutuje też artykuł w *Postępach Fizyki* dotyczący nagrody Nobla z fizyki 2024 [4].

stwardnienie zanikowe boczne, które zabija neurony niosące z mózgu informację sterującą mięśniami. Okrucieństwo tej choroby polega na tym, że sam mózg pozostaje względnie nietknięty, ale już na zawsze w piekle zamknięcia (ang. *locked-in state*), gdyż cała czynna komunikacja jest przez ludzi przekazywana za pośrednictwem mięśni: płuc, krtani, twarzy czy rąk. Kiedy sterowanie wszystkimi tymi mięśniami zostaje przerwane, tracimy możliwość wyrażenia czegokolwiek. Chyba żeby intencję dało się odczytać bezpośrednio z mózgu. I to jest właśnie klasyczna definicja BCI: odczyt generowanych w mózgu intencji bez pośrednictwa mięśni.



Rys. 6. Pierwszy w Polsce publiczny pokaz działania interfejsu mózg-komputer, czerwiec 2008, WF UW, Hoża 69

Jak je odczytać? Jak wspomniano w rozdziale 2, przetwarzanie informacji mózgu wiąże się z powstawaniem potencjałów elektrycznych. Ich ślady, czyli elektroencefalogram (EEG), odczytujemy z elektrod umieszczonych na powierzchni głowy od niemal stulecia [3] – do dzisiaj EEG jest najpopularniejszą z technik rejestracji procesów zachodzących w mózgu. Podczas pierwszego w Polsce publicznego pokazu BCI wykorzystano EEG (rys. 6): na zdjęciu widać obraz oglądany przeze mnie wówczas na ekranie laptopa, na którym migają kolejno wiersze i kolumny macierzy symboli. Zadaniem systemu jest wykrycie, na którym znaku koncentruję uwagę – reakcja nastąpi, gdy jednocześnie zostaną podświetlone: odpowiedni wiersz i odpowiednia kolumna. Reakcja ta ma być oczywiście wykrywana bezpośrednio z elektrycznych śladów myśli, czyli z EEG. Ta reakcja to tak zwany potencjał wywołany – zjawisko znane w encefalografii od dziesięcioleci: potencjał (załamek), widoczny po uśrednieniu kilku czy kilkuset odcinków EEG, zsynchronizowanych z bodźcem. Bodźce generuje komputer, takie uśrednianie może więc zachodzić w czasie prawie rzeczywistym, jak to widać na rys. 7, gdzie zielone krzywe odpowiadają średnim odcinków zsynchronizowanych z bodźcem, na który użytkownik miał zwracać uwagę (TGT od ang. *target*), a niebieskie – pozostałych. Zielone krzywe wykazują odchylenia w okolicy 300 ms po bodźcu, zwane potencjałem (załamkiem) P300. Jest to potencjał uwagowy, wywoływany bodźcami, na które zwracamy uwagę, czyli zależny od naszej woli. Dzięki temu możemy go wykorzystać do sterowania i przekazywania informacji. Jeśli chcemy powiedzieć TAK, koncentrujemy uwagę na wystąpieniach odpowiednio oznaczonego bodźca, na przykład uważnie licząc jego mignięcia. Jeśli system poprawnie przypisze potencjał do bodźca, na którym koncentrowaliśmy uwagę, przekażemy bez pośrednictwa mięśni co najmniej jeden bit.



Rys. 7. Panel kontroli BCI w czasie rzeczywistym. W górnej części panelu po lewej: sygnał EEG mierzony z elektrod, których symbole na głowie widocznej po prawej pokazują aktualne oporności. W lewym dolnym rogu panelu: widzimy potencjały dla każdej elektrody, uśredniane z napływających danych, synchronizowane z mignięciami obu kwadratów, a obok rozkład prawdopodobieństw klasyfikatora, który będzie rozpoznawał wystąpienie potencjału na zielonych bądź niebieskich krzywych, odczytując w ten sposób intencje użytkownika (wybór TAK/NIE) bez pośrednictwa mięśni. W prawym dolnym rogu rysunku: zdjęcie ekranu, na którym migają kwadraty TAK i NIE.

Tak właśnie działają współczesne BCI: z mierzonych różnymi metodami śladów aktywności mózgu próbują odczytać (sklasyfikować) intencje, którym możemy przypisać mniej czy bardziej umowne znaczenie. To znaczy, że aby zamówić przez BCI kawę, nie wystarczy pomyśleć o pachnącej filiżance – trzeba skonstruować interfejs, w którym będzie opcja wyboru kawy i przypisać jej mierzalną reakcję, którą użytkownik może kontrolować, na przykład wspomniany potencjał P300.

BCI działają znacznie efektywniej, jeśli zaczniemy od sesji kalibracyjnej: użytkownik proszony jest o koncentrację uwagi na wskazywanych przez system bodźcach, dzięki czemu możemy wykreślić widoczne na rys. 7 krzywe i zoptymalizować klasyfikator. W klasycznym podejściu wystarczy zwykle kilka-kilkanaście minut takiej kalibracji, aby osiągnąć realną szybkość działania rzędu kilku-kilkunastu liter na minutę.

Co innego, jeśli chcemy bić rekordy szybkości. Formuła 1 w BCI to rejestracje potencjałów z wnętrza czaszki (elektrokortykografia) i klasyfikatory wykorzystujące głębokie sieci neuronowe. Aktualny rekord zapisany w księdze Guinessa (78 słów na minutę przy 25% błędów) osiągnięto na sygnałach z 253 elektrod umieszczonych bezpośrednio na korze mózgowej, po tygodniach sesji kalibracyjnych, w czasie których sieć uczyła się dopasowywać odczytywane z elektrod wzorce aktywności mózgowej do wypowiadanych bezgłośnie przez pacjentkę słów z wybranego zestawu [17]. Ale wciąż jest to tylko dopasowywanie wzorców neuronalnej aktywności, odpowiadających za przygotowanie ruchów warg i języka, do wypowiadanych bezgłośnie słów, a nie odczytywanie myśli.

Fakt, że BCI nie odczytują myśli, tylko świadomie generowane intencje i to wyłącznie w bardzo specyficznych sytuacjach, wymagających uprzedniej świadomej współpracy pacjenta[6] w procesie kalibracji, nie wpływał na liczbę alarmistycznych wizji „podsłuchiwania myśli" i „końca prywatności", jakie wypełniały popularne media w czasach szczytu popularności BCI ok. dekadę temu. Nie przypominamy tego faktu, żeby ośmieszyć modne dzisiaj dyskusje o zagrożeniach AI, gdyż te zagrożenia są realne i mamy z nimi do czynienia już teraz. Niestety zagrożenia, o których dyskutują celebryci w mediach, są zwykle bardziej futurystyczne i medialne, niż aktualne i konkretne. Realnymi i konkretnymi zagrożeniami związanymi z BCI zajmuje się neuroetyka, a do zagrożeń AI wrócimy w następnym rozdziale.

6. Potencjał P300 można zwykle wykryć również bez współpracy badanego, jednak działa to znacznie gorzej niż po kalibracji, dlatego wykorzystujące P300 detektory kłamstw działają niewiele lepiej od klasycznych wariografów.

Warto też zwrócić uwagę na afiliacje autorów pracy [17]: nie jest to Neuralink, tylko Uniwersytet Kalifornii. A jednak Internet „rozgrzewają do czerwoności" niemal wyłącznie doniesienia o kolejnym (trzecim?) pacjencie, który porusza kursorem za pośrednictwem interfejsu Neuralink. Pomijany jest przy tym fakt, że od czasu publikacji [10] w roku 2006, czyli na długo przed powstaniem Neuralink, w licznych ośrodkach naukowych wszczepiono już podobne interfejsy kilkudziesięciu pacjentom, którzy założyli nawet „koalicję pionierów BCI" https://bcipioneers.org. Neuralink nie wniósł tu nic nowego, poza deklarowanym usprawnieniem samego procesu wszczepiania implantu. Czyż to nie za mało dla inwestorów, dzięki którym firma jest wyceniana na 8 miliardów dolarów...? I może właśnie dlatego, zgodnie z kultowym w Krzemowej Dolinie aforyzmem *fake it till you make it*, Elon Musk obiecuje zrewolucjonizowanie leczenia choroby Parkinsona, epilepsji, autyzmu, otyłości, depresji, schizofrenii... W tym kontekście należy też oceniać obietnice bezpośredniego połączenia mózgu z AI.

5. Apokalipsa AI

Skoro nie widać bliskich perspektyw na połączenie naszych mózgów z AI, ani na przeniesienie świadomości do cyberprzestrzeni (rozdział 2), pozostaje bliżej przyjrzeć się zagrożeniom, jakie niesie dla ludzkości gwałtowny rozwój tych technologii. Mówią o nich wszyscy – od youtuberów do noblistów z dziedziny, jak Geoffrey Hinton i Demis Hassabis. Tylko zwykle dość ogólnikowo.

Większość apokaliptycznych przepowiedni wiązana jest z oczekiwanym nadejściem silnej (ogólnej) AI (ang. *artificial general intelligence*, AGI), która ma już niedługo przewyższyć inteligencję ludzką pod każdym względem. Gdy tylko uzyska sprawczość, stanie się jasne, „kto tu rządzi".[7]

Ostatnie badania wydają się też potwierdzać tezę, że do tego punktu zbliżamy się z dwu stron: o ile LLM, dzięki konwersacjom z ludźmi, gromadzą coraz więcej danych, to ludzie korzystający na co dzień z narzędzi AI wydają się zatracać zdolności krytycznego myślenia [14].

Z kolei według scenariuszy rodem z *science-fiction*, apokalipsa AI może wyniknąć z nieporozumienia. Szwedzki filozof Nick Bostrom zaproponował eksperyment myślowy, w którym zarządzanie fabryką spinaczy biurowych oddajemy całkowicie w ręce AI, pozostawiając jako jedyny cel maksymalizację produkcji. AI słusznie uznaje, że ludzie mogą w tym procesie tylko przeszkadzać i przerabia ich na spinacze.

I tak dalej. Lubimy słuchać takich przepowiedni, gdyż (1) odnoszą się do przyszłości, więc dla większości wydają

7. Przyjmuje się, że światem rządzą ludzie, nie ameby, właśnie z powodu różnic w poziomie inteligencji.

się, niestety, równie niegroźne jak globalne ocieplenie i (2) niejako automatycznie zwalniają nas z myślenia – przecież i tak nic nie poradzimy w obliczu wszechmocnej AGI.

Skoro jednak, Drogi Czytelniku, dotarłeś niemal do końca tego eseju, to mam nadzieję, że docenisz próbę analizy *realnych* zagrożeń i szkód, jakie zaliczane ostatnio do AI algorytmy powodują od dziesięcioleci.

5.1. Rząd dusz

Głównym miernikiem wartości i źródłem ogromnych dochodów platform mediów społecznościowych jest liczba aktywnych użytkowników i czas przez nich spędzany na przeglądaniu treści podsuwanych przez serwisy. Idea nie jest nowa, ponieważ media zawsze walczyły o uwagę użytkowników różnymi sposobami: od taniej sensacji do dziennikarskiej rzetelności. Sytuacja mediów społecznościowych jest inna o tyle, że autorami treści są w większości użytkownicy. Daje to pretekst do tyleż wygodnego, co nieetycznego zrzekania się odpowiedzialności przez właścicieli platform. Szczególnie dlatego, że wybór treści podsuwanych użytkownikom mediów społecznościowych „na pierwszej stronie", czyli decyzje o tym, które posty są wzmacniane i promowane, podejmują algorytmy.

Algorytmy te, wedle dzisiejszych definicji (rozdział 1) uznawane za AI, dość szybko „odkryły", że największe zaangażowanie użytkowników gwarantują treści brutalne i antagonizujące, a ich ewentualna prawdziwość ma na zaangażowanie wpływ co najwyżej marginalny. Prawda jest zwykle mniej ciekawa i trudniejsza do zrozumienia od wymyślanych opowieści. Prowadzi to do propagowania i wzmacniania treści bardzo często szkodliwych i fałszywych. Antagonizowanie przeciwko sobie grup społecznych, etniczych i całych narodów, nie jest w tym przypadku częścią tajnego planu, chodzi bowiem tylko o maksymalizowanie zysku firm, które wciąż unikają odpowiedzialności za konsekwencje.

A tragiczne konsekwencje w tym przypadku nie są już hipotetyczne, tylko konkretne i udokumentowane. Jak na przykład ludobójstwo i czystki etniczne w Myanmar (dawniej Birma) w latach 2016–2017 [6], wynikłe w dużej części z rozpropagowania za pośrednictwem platformy Facebook mowy nienawiści ultranacjonalistycznego mnicha buddyjskiego Ashina Wirathu, którego „atrakcyjne" posty szkalujące muzułmańską grupę etniczną Rohingya były przez algorytmy powielane i propagowane, w przeciwieństwie do „nudnych" opinii wielu innych mnichów wzywających do współczucia. Empatia okazała się mniej angażująca od nawoływania do przemocy – nie przykuwała uwagi użytkowników Facebooka.

Efektem, który niejako przy okazji wywołują algorytmy rekomendujące treści, jest błąd potwierdzenia (ang. *confirmation bias*): w serwisie widzimy tylko treści odpowiadające naszym przekonaniom i przesądom i stajemy się coraz bardziej odizolowani od argumentów przeciwnych. W ten sposób AI doprowadza do polaryzacji grup społecznych funkcjonujących w odrębnych bańkach informacyjnych.

5.2. Mikrotargetowanie

Mikrotargetowanie (ang. *microtargetting*) to kolejny, z pozoru niewinny, mechanizm, zwiększający efektywność reklam. Reklamy mają nas zainteresować konkretnymi produktami, ale nie wszyscy interesują się tym samym i nie na wszystkich działają takie same argumenty. Skąd algorytmy wiedzą, jak przekonywać konkretne osoby? Michał Kosiński pokazał, że opatentowany przez Facebook algorytm (patent US20160283485A1) [13]:

jest w stanie określić preferencje seksualne (u mężczyzn skutecznie w 88% przypadków), wygląd, zainteresowania, poziom inteligencji, pochodzenie etniczne i kolor skóry (u Amerykanów skutecznie w 95% przypadków), wyznanie, poziom zadowolenia z życia, uzależnienia, wiek, płeć oraz poglądy społeczne, religijne i polityczne […] *na podstawie 68 polubień na Facebooku.*

Zastosowanie technik manipulacji behawioralnej, celowanej precyzyjnie w indywidualne lęki i słabości każdego z nas, daje niemal nieograniczone możliwości kształtowania opinii, w tym wpływania na wyniki wyborów i referendów, których obiektywność stanowi fundament demokracji. Przykładem może być domniemany wpływ firmy Cambridge Analitica na wybory w USA i Brexit.

5.3. Nieznośna lekkość fałszowania

To już historia najnowsza, pisana przez generatywną AI. Nowością nie są fałszerstwa jako takie, tylko niemal nieograniczona dostępność niemal doskonałych narzędzi, umożliwiających tworzenie z pomocą AI niemal doskonałych fałszerstw – praktycznie dla każdego, bez wielkich nakładów czy specjalistycznej wiedzy. Komunikacja między ludźmi polega dzisiaj głównie na cyfrowej wymianie informacji. Jej wiarygodność to filar demokracji. Strzec jej powinny państwa tak, jak strzegą wiarygodności pieniądza jako umowy społecznej. Fałszerstwa banknotów są rzadkie nie z przyczyn technicznych, tylko ze względu na surowe w tym zakresie prawo. Miejmy nadzieję, że wspomniane na początku Rozporządzenie PE i Rady UE [21] zadziała podobnie przynajmniej w Europie. Ale to już inna historia.

Niniejszy esej powstał na podstawie materiałów do wykładu dla studentów Wydziału Fizyki UW https://brain.fuw.edu.pl/edu/index.php/Technologie_informacyjne_i_komunikacyjne.

Literatura

[1] DeepSeek-AI et al. *Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. 2025. arXiv: 2501.12948

[2] Emily M. Bender i in. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" W: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA, 2021, s. 610–623. DOI: 10.1145/3442188.3445922.

[3] Hans Berger. "Über das Elektrenkephalogramm des Menschen". W: *Archiv für Psychiatrie und Nervenkrankheiten* 87.1 (1929), s. 527–570. DOI: 10.1007/BF01797193.

[4] Anna Dawid. "Ktokolwiek widział, ktokolwiek wie! Ukradziono Nagrodę Nobla z fizyki!" W: Postępy Fizyki 3–4 (75 2024). DOI: 10.61947/uw.PF.2024.75.3-4.12-16.

[5] Jia Deng i in. "ImageNet: A large-scale hierarchical image database". W: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, s. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[6] Christina Fink. "Dangerous speech, antimuslim violence, and Facebook in Myanmar". W: *Journal of International Affairs* 71.1.5 (2018), s. 43–52. ISSN: 0022197X. URL: https://www.jstor.org/stable/26508117 (dostęp 12. 02. 2025).

[7] R. A. Fisher. "The use of multiple measu- rements in taxonomic problems". W: *Annalsof Eugenics* 7.2 (wrz. 1936), s. 179–188. ISSN: 2050-1420 (print), 2050-1439 (electronic). DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[8] Ian Goodfellow, Yoshua Bengio i Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[9] Michael Townsen Hicks, James Humphries i Joe Slater. "ChatGPT is bullshit". W: *Ethics and Information Technology* 26.2 (2024), s. 38. DOI: 10.1007/s10676-024-09775-5.

[10] Leigh R. Hochberg i in. "Neuronal ensemble control of prosthetic devices by a human with tetraplegia". W: *Nature* 442.7099 (2006), s. 164–171. DOI: 10.1038/nature04970.

[11] A. L. Hodgkin and A. F. Huxley. "A quantitative description of membrane current and its appli-cation to conduction and excitation in nerve". W: *The Journal of Physiology* 117.4 (1952), s. 500–544. DOI: 10.1113/jphysiol.1952.sp004764.

[12] John Jumper i in. "Highly accurate protein structure prediction with AlphaFold". W: *Nature* 596.7873 (2021), s. 583–589. DOI: 10.1038/s41586-021-03819-2.

[13] Michał Kosiński, David Stillwell i Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior". W: *Proceedings of the National Academy of Sciences* 110.15 (2013), s. 5802–5805. DOI: 10.1073/pnas.1218772110.

[14] Hao-Ping (Hank) Lee i in. "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers". W: *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, kw. 2025. URL: https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/.

[15] Henry Markram. *A brain in a supercomputer [video]*. https://www.ted.com/talks/henry_markram_a_brain_in_a_supercomputer, Lip. 2009.

[16] Warren McCulloch and Walter Pitts. "A logical calculus of ideas immanent in nervous activity". W: *Bulletin of Mathematical Biophysics* 5 (1943), s. 127–147.

[17] Sean L. Metzger i in. "A high-performance neuroprosthesis for speech decoding and avatar control". W: *Nature* 620.7976 (2023), s. 1037–1046. DOI: 10.1038/s41586-023-06443-4.

[18] Adam Mickiewicz. *Dziadzy część III*. https://polona.pl/preview/626a95ba-f7ea-4082-98a8-2a6648dd65c4. Dostęp: 2025.02.08. Paryż, 1838.

[19] Iman Mirzadeh i in. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. 2024. arXiv: 2410.05229

[20] H.P. Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988. isbn: 9780674576162.

[21] Rada Unii Europejskiej Parlament Europejski. Rozporządzenie parlamentu europejskiego i rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji) (tekst mający znaczenie dla EOG). https://eur-lex.europa.eu/legal-content/PL/ALL/?uri=CELEX:32024R1689, 2024.

[22] *Philosophical transactions of the royal society B 373 (1758 2018): Connectome to behaviour: modelling C. elegans at cellular resolution. Discussion meeting issue organized and edited by Stephen D. Larson, Padraig Gleeson and André E.X. Brown*. URL: https://royalsocietypublishing.org/toc/rstb/2018/373/1758.

[23] Jiawei Su, Danilo Vasconcellos Vargas i Kouichi Sakurai. "One Pixel Attack for Fooling Deep Neural Networks". W: *IEEE Transactions on Evolutionary Computation* 23.5 (paź. 2019), s. 828–841. ISSN: 1941-0026. DOI: 10.1109/tevc.2019.2890858.

[24] Min-Jen Tsai, Ping-Yi Lin i Ming-En Lee. "Adversarial Attacks on Medical Image Classification". W: *Cancers* 15.17 (2023). ISSN: 2072-6694. DOI: 10.3390/cancers15174228.

[25] Ashish Vaswani i in. "Attention is All you Need". W: *Advances in Neural Information Processing Systems*. Red. I. Guyon i in. T. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[26] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". W: *Commun. ACM* 9.1 (sty. 1966), s. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168.

[27] John Graham White i in. "The Structure of the Nervous System of the Nematode Caenorhabditis elegans". W: *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 314.1165 (1986), s. 1–340. ISSN: 0962-8436. URL: https://royalsocietypublishing.org/toc/rstb/2018/373/1758.