Sztuczna inteligencja – jej rozwój, szanse i zagrożenia

prof. dr hab. inż. Roman Słowiński
ORCID: 0000-0002-5200-7795
Instytut Informatyki Politechniki Poznańskiej i Instytut Badań Systemowych PAN

1. Wstęp

Od blisko 80 lat obserwujemy niezwykły postęp w technice obliczeń cyfrowych. Technika komputerowa wkracza do wszystkich dziedzin naszego życia i oferuje coraz to nowe usługi w rozwiązywaniu zadań intelektualnych, które człowiek wykonałby mniej efektywnie. Rozwija się sztuczna inteligencja (SI), która ma ambicje przekroczenia ludzkich zdolności świadomego tworzenia. Stawiamy wobec tego pytania o świadomość inteligentnych maszyn i możliwość ich „uczłowieczenia", a nawet przekroczenia ograniczeń ludzkiej biologii.

W niniejszym wykładzie przedstawimy ewolucję rozumienia sztucznej inteligencji oraz szanse i zagrożenia związane z jej dalszym rozwojem. Prześledzimy trzy następujące po sobie w czasie interpretacje pojęcia inteligencji maszyn i powiązane z nimi efekty. Według nich inteligencja maszyn jest: (i) symulacją procesów myślowych człowieka rozumianych jako obliczanie (arytmetyczne i logiczne), (ii) wykonywaniem zadań intelektualnych człowieka (klasyfikacja, tworzenie reguł, odkrywanie praw i relacji), (iii) racjonalnym działaniem motywowanym planem i wiedzą o otoczeniu.

2. Inteligencja maszyn jako symulacja procesów myślowych człowieka będących obliczaniem

Pogląd, że inteligencja człowieka może być symulowana przez maszyny jako proces obliczania, panował, zanim jeszcze pojawiło się pojęcie SI. Pojawiło się ono po raz pierwszy w 1956 r. podczas *Dartmouth Summer Research Project on Artificial Intelligence* – konferencji zorganizowanej w Dartmouth College w Hanowerze, USA.

W 1950 r. Alan Turing, ojciec nowoczesnej nauki o komputerach, w czasopiśmie „Mind”[1] postawił hipotezę, że myślenie jest obliczaniem i zapytał „czy maszyny mogą myśleć?”. Jeśli wszelkie myślenie byłoby obliczaniem, to można by przypisać procedurze algorytmicznej moc wytwarzania świadomości. To by z kolei oznaczało, że maszyna cyfrowa może przerosnąć inteligencję człowieka po prostu przez odpowiednie zwiększenie efektywności obliczeń.

Powyższe pytanie stawiali też logicy i matematycy przed powstaniem koncepcji współczesnej maszyny cyfrowej. Wielki matematyk XIX w. Georg Cantor uważał, że „nowoczesna matematyka posługując się pojęciem nieskończoności wychodzi poza świat materialny”. Mówił, że „przyjęcie nieskończoności, to jak wiara w Boga”. Człowiek ma zdolność posługiwania się pojęciem nieskończoności, podczas gdy maszyna jej nie ma. Jednakże, jako argument o wyższości umysłu ludzkiego nad maszyną jest dość słaby, gdyż w praktyce wystarczają liczby skończone.

Nieco później nadzieją na matematyczny dowód wyższości umysłu nad maszyną były między innymi prace amerykańskiego logika polskiego pochodzenia Emila Leona Posta i austriackiego logika Kurta Gödla[2]. Post był jednym z pierwszych, którzy zwrócili uwagę na ograniczenia maszyn obliczeniowych i granice ludzkiej wiedzy w kontekście rozwiązywania problemów matematycznych. Gödel natomiast znany jest z twierdzenia o niezupełności systemów formalnych opartych na arytmetyce, które mówi, że „nie istnieje algorytm, który przy danym układzie aksjomatów podałby dowód każdego prawdziwego twierdzenia arytmetyki”. Według brytyjskiego filozofa Johna Randolpha Lucasa[3] przemawiałoby ono za wyższością umysłu nad maszyną, gdyż niealgorytmiczny umysł nie mógłby być adekwatnie symulowany przez komputer. Ludzie, którzy jednak nie do końca rozumieją pojęcie liczb (np. nie potrafimy udowodnić hipotezy Riemanna), nie mogą przekazać tego pojęcia maszynie, czyli nie mogą zadać jej wszystkich możliwych aksjomatów[4]. Niezależnie od tego, że jak pisze Stanisław Krajewski[5], argument Lucasa zawiera sprzeczność z powodu autoreferencji, bardziej wnikliwa analiza twierdzenia Gödla, potwierdzona przez niego samego, pokazuje, że jego twierdzenie nie wyklucza

[1] A.M. Turing, Computing machinery and intelligence, *Mind*, 59, 236 (1950) 433–460.

[2] Ciekawą analizę historyczną tego procesu znajdziesz w wykładzie prof. Giovanniego Sommarugi z ETH Zürich, https://www.academia.edu/42767594/

[3] J.R. Lucas, Minds, Machines and Gödel, *Philosophy*, 36 (1961).

[4] St. Krajewski, *Twierdzenie Gödla i jego filozoficzne interpretacje. Od mechanicyzmu do postmodernizmu*, Wydawnictwo IFiS PAN, Warszawa, 2003.

[5] St. Krajewski, On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to 'Out-Gödel' the Mechanist, *Fundamenta Informaticae*, 81, 1–3 (2007) 173–181.

istnienia formalnej teorii, która obejmowałaby całą naszą matematykę, to znaczy, że wyniki Gödla nie wykluczają istnienia robota równoważnego umysłowi w zakresie matematyki. Jednakże, jak zauważa Krajewski: „gdyby taka teoria lub robot istniały, nie bylibyśmy w stanie pojąć ani ich równoważności z nami, ani nawet ich niesprzeczności czy poprawności. Sam wynik Gödla nie rozwiązuje filozoficznych debat; nie dowodzi ani antymechanicyzmu, ani realizmu, choć dostarcza pewnych wskazówek. Praca Gödla podważa natomiast jedną z radykalnych ideologii SI, tzw. «silną SI» Searle'a. Można dowieść, że nie jesteśmy w stanie zaprojektować programu ani robota równoważnego ludzkiemu umysłowi, nawet jeśli taki program teoretycznie mógłby istnieć. Twierdzenie Gödla dopuszcza jednak możliwość, że taki program lub robot mógłby powstać w wyniku działania jakiejś superinteligencji lub w wyniku ewolucji w społeczeństwie robotów. W takim przypadku nie bylibyśmy jednak w stanie odkryć ani potwierdzić tej równoważności"[6].

Pomimo tej nierozstrzygniętej debaty na temat wyższości inteligencji ludzkiej nad maszynową, nie ulegało wątpliwości, że maszyny osiągną pewien poziom inteligencji i aby porównać ją z inteligencją człowieka Alan Turing zaproponował dobrze znaną grę imitacyjną, zwaną testem Turinga. Polega on na tym, że maszyna udzielająca odpowiedzi na pytania operatora powinna sprawiać wrażenie, że odpowiada człowiek. Aby zdać test Turinga, maszyna musi być zdolna do przetwarzania języka naturalnego, rozpoznawania i syntezy mowy, przetwarzania i rozpoznawania obrazów, odkrywania wiedzy z danych (dziś z Internetu) i automatycznego wnioskowania.

Oczywiście, w latach pięćdziesiątych ubiegłego wieku trudno było wyobrazić sobie taką maszynę, ale stopniowo komputery nabywały te umiejętności. Pierwszym spektakularnym zwycięstwem maszyny nad człowiekiem była wygrana komputera „Deep Blue" firmy IBM z arcymistrzem szachowym Gari Kasparowem w 1997 r. W 2014 r. program, który „udawał" 13-letniego Eugene'a Goostmana po kilkuminutowym czacie przekonał 30% sędziów, że jest człowiekiem, czyli zdał w tym stopniu test Turinga. Program ten umiał dość dobrze rozumieć i formułować zdania w języku naturalnym, by posługując się wiedzą z Internetu, odpowiadać „rozsądnie" na pytania. To prawda, że Eugene tłumaczył swoją nie najlepszą angielszczyznę pochodzeniem ukraińskim i wykręcał się od wielu pytań młodzieżowymi odzywkami typu „guzik mnie to obchodzi" albo „mam to gdzieś", ale dla części sędziów robił to jak autentyczny nastolatek. Dwa lata później w Seulu system „AlphaGo" firmy Google pokonał mistrza gry w „Go" Lee Sedola. W grze w szachy komputer już nie mierzy się z człowiekiem, lecz z innymi programami. W 2017 r. firma DeepMind pokazała system „AlphaZero" do gry w szachy, oparty

[6] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

na sztucznej sieci neuronowej, który pobił „Stockfish", najlepszy dotąd system ogrywający arcymistrzów. W tym samym roku do mediów trafiła wiadomość, że król Arabii Saudyjskiej przyznał obywatelstwo swojego kraju robotowi „Sophia" firmy Hanston Robotics z Hongkongu. Niedługo potem, w 2018 r., IBM Watson Debater wygrał debatę oksfordzką, a wcześniej turniej Va-Banque. Dzisiaj duże modele językowe, jak ChatGPT, znacznie lepiej od Goostmana czy Sophii odpowiadają na pytania i mają już miliardy użytkowników.

Dochodzimy do wniosku, że test Turinga jest przestarzały, gdyż testowanie wiedzy i logicznego wnioskowania nie rozstrzyga o równości inteligencji ludzkiej i maszynowej. Ciekawą analizę poglądów na temat testu Turinga przeprowadzili Józef Bremer i Mariusz Flasiński[7]. Porównali poglądy czterech wpływowych reprezentantów filozofii analitycznej: Ludwiga Wittgensteina, Noama Chomsky'ego, Hilarego Putnama i Johna Searle'a. Na podstawie różnych argumentów dotyczących kwestii filozoficznych i/lub językowych dochodzą oni do odmiennych ocen zarówno znaczenia, jak i przydatności testu Turinga dla rozwoju nauk kognitywnych i modeli SI. Niemniej jednak łączy ich odrzucenie idei, że test Turinga można traktować jako test „maszynowego myślenia". Wydaje się, że wynika to z troski o właściwe użycie języka – kwestii będącej fundamentalnym zobowiązaniem metodologicznym filozofii analitycznej.

Wspomniany już Stanisław Krajewski[8] zaproponował z kolei wzmocnienie testu Turinga, czyli eksperyment myślowy, w którym dziecko jest od urodzenia wychowywane przez roboty i ma się okazać, czy tak ukształtowany człowiek będzie w miarę „normalny". Autor dodaje od razu, że taki test byłby oczywiście nieetyczny.

3. Inteligencja maszyn jako zdolność do wykonywania intelektualnych zadań człowieka

Początkowa euforia związana ze sztuczną inteligencją, która sprawi, że wkrótce maszyna zastąpi człowieka w myśleniu, ustąpiła w latach sześćdziesiątych realizmowi spójnemu ze skromniejszą definicją sztucznej inteligencji podaną przez Norberta Wienera: „inteligencja jest procesem pozyskiwania i przetwarzania informacji dla osiągania wyznaczonych celów". Tak rozumiana sztuczna inteligencja odnosi spektakularne sukcesy w przetwarzaniu i rozpoznawaniu obrazów, rozpoznawaniu

[7] J. Bremer, M. Flasiński, The Turing Test, or a Misuse of Language when Ascribing Mental Qualities to Machines. *Forum Philosophicum*, 27, 1 (2022) 6–25.

[8] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

i syntezie mowy oraz analizie tekstów i przetwarzaniu języka naturalnego. Inteligentne wykonywanie zadań odbywa się z użyciem heurystyk, które pozwalają ograniczyć zakres przeszukiwania pola problemowego, a tym samym skrócić czas rozwiązywania problemu. Ponadto temu procesowi towarzyszy uczenie się. Jest to rozwój pozytywny, sprzyjający symbiozie pomiędzy ludźmi i maszynami w systemach inteligentnego wspomagania decyzji[9].

Chociaż postępy w wykonywaniu zadań intelektualnych człowieka przez maszyny nie zawsze są tak spektakularne, jak pokonanie szachowego arcymistrza, trzeba przyznać, że na polu eksploracji dużych wolumenów danych (ang. *big data*) człowiek został pokonany. Od początku XXI w. rozwija się analityka danych (ang. *data-driven analytics*) polegająca na komputerowym przekształcaniu surowych danych pochodzących z różnych źródeł w wiedzę wykorzystywaną do podejmowania lepszych decyzji. Do obiegu weszło także określenie nauka o danych (ang. *data science*).

Na zasadzie indukcji odkrywa się wzorce prawdziwe w świecie analizowanych danych, które reprezentują wiedzę zawartą w tych danych. Jest to paradygmat tzw. uczenia maszynowego (ang. *machine learning*). W 1983 r. Herbert Simon zdefiniował uczenie maszynowe jako „adaptacyjne zmiany w systemie, które pozwalają systemowi wykonać za następnym razem takie samo zadanie lub zadania podobne bardziej efektywnie". Odkrywanie wiedzy z danych odbywa się w sposób nieukierunkowany: np. „powiedz mi coś interesującego o moich danych"; ukierunkowany: np. „scharakteryzuj klientów, którzy odpowiedzieli na ofertę promocyjną"; przez testowanie lub uściślanie hipotez: np. „czy to prawda, że jest związek między sposobem odżywiania a zapadalnością na chorobę X?"

Warto dodać, że w coraz większym stopniu analitycy danych radzą sobie z danymi niepełnymi, niedokładnymi, podlegającymi przypadkowym fluktuacjom i częściowo sprzecznymi – te niedoskonałości danych są problemem we wnioskowaniu automatycznym pomimo wielkiej masy danych.

W przypadku danych niepełnych i sprzecznych na uwagę zasługuje matematyczna teoria polskiego informatyka Zdzisława Pawlaka oparta na pojęciu zbioru przybliżonego (ang. *rough set*)[10]. Warto zauważyć, że John Hopfield otrzymał w 2024 r. Nagrodę Nobla za opracowanie neuronowej sieci asocjacyjnej i prace, które przyczyniły się do rozwoju metod pozwalających na wnioskowanie z niepełnych danych właśnie.

---

[9] R. Słowiński (ed.): *Intelligent Decision Support – Handbook of Applications and Advances of Rough Sets Theory*. Kluwer Academic Publishers, Dordrecht, 1992.

[10] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1991.

W tym kontekście nie mogę się także powstrzymać od wspomnienia o Dominacyjnej Teorii Zbiorów Przybliżonych (ang. *Dominance-based Rough Set Approach*), którą zaproponowałem z kolegami z Katanii[11]. To podejście dokonało swoistego przełomu w rozumowaniu na temat danych porządkowych, typowych dla wieloatrybutowych problemów decyzyjnych. Radzi sobie ono z danymi niepełnymi i częściowo sprzecznymi, i zapewnia reprezentację preferencji użytkownika w postaci zrozumiałych reguł decyzyjnych o składni „*jeśli...*, *to...*". Na przykład, we wspomaganiu decyzji w warunkach ryzyka i niepewności wspomaganie decyzji dokonuje się za pomocą reguł typu „*jeśli prawdopodobieństwo zysku co najmniej 200$ więcej jest* ≥ 0.75, *a zysk co najmniej 100$ więcej jest pewny*, *to akt inwestycyjny A jest lepszy od aktu B*", zamiast za pomocą argumentu teorii użyteczności, który stwierdza, że „*akt A jest lepszy od aktu B*, *gdyż spodziewana użyteczność* $U(A) = 0.67 > U(B) = 0.63$", nie wyjaśniając, co kryje się pod tymi liczbami. Podejście regułowe do wspomagania decyzji jest zgodne z wyjaśniającymi ambicjami sztucznej inteligencji (ang. *explainable AI* lub *XAI*).

Wobec rosnącej efektywności maszyn cyfrowych w odkrywaniu wiedzy z dużych wolumenów danych pojawia się często pytanie, jaką prawdę odkrywa maszyna, czy jest to ta sama prawda, jakiej szuka człowiek? Otóż maszyna pozna prawdę zawartą w tych danych, z których została wyindukowana. Mogą jednak pojawić się nowe dane, które starą prawdę zafałszują. To pokazuje ograniczoność pojęcia prawdy odkrytej przez indukcję charakterystyczną dla uczenia maszynowego. Zbiór danych, z których uczy się maszyna, nie jest ani wyczerpujący, ani stały. Prawda, którą kontempluje człowiek, została dobrze scharakteryzowana przez św. Jana Pawła II w encyklice *Fides et Ratio*[12]. Do zrozumienia tej prawdy potrzeba nie tylko racjonalności, ale także wiary w tajemnicę stworzenia. Jak pisze Eric-Emmanuel Schmitt: „tajemnica nie tkwi w tym, co nieznane, lecz w tym, co niezrozumiałe"[13]. Gdyby cała prawda o Bogu i człowieku była osiągalna tylko „szkiełkiem i okiem", to nie bylibyśmy osobami wolnymi i nie moglibyśmy wybierać – czy wierzę, czy nie wierzę. Akceptuję tajemnicę, niepewność istnienia Boga, sercem, ale wiarę przyjmuję racjonalnie z własnej woli[14]. Tak dochodzę do prawdy, której nie mogą zafałszować nowe dane.

---

[11] S. Greco, B. Matarazzo, R. Słowiński, Rough sets theory for multicriteria decision analysis. *European J. of Operational Research*, 129 (2001) 1–47.

[12] https://www.vatican.va/content/john-paul-ii/pl/encyclicals/documents/hf_jp-ii_enc_14091998_fides-et-ratio.html

[13] E.-E. Schmitt, *Sen o Jerozolimie*. Znak Literanova, Kraków, 2024.

[14] Rozmowa z prof. R. Słowińskim, Człowiek – robot, robot – człowiek? *W drodze*, 3 (2020) 66–77.

Omawiając rozwój zdolności maszyn cyfrowych do wykonywania zadań intelektualnych człowieka, trudno jest pominąć rozwój sieci Internet, która jest ogromną bazą danych. Dane te są przedmiotem analiz algorytmów uczenia maszynowego. Dzisiejszy Internet Rzeczy (ang. *Internet of Things – IoT*) łączy ponad 75 miliardów urządzeń identyfikowanych przez adresy IP. Są to nie tylko komputery, ale liczne urządzenia typu telefony, czujniki, pojazdy, kamery, a nawet lodówki. Internet sprawił, że świat stał się „globalną wioską", w której wszyscy mogą się komunikować ze wszystkimi, oczywiście pod warunkiem posiadania dostępu do sieci (brak dostępu nazywa się dzisiaj „cyfrowym wykluczeniem").

Tak wielki przyrost liczby urządzeń podłączonych do sieci i generujących dane, a także postęp algorytmów uczenia maszynowego nie jest oczywiście obojętny ani dla życia społecznego, ani dla nauki i jej przestrzennej organizacji. Z punktu widzenia nauki sieć komputerowa i algorytmy sztucznej inteligencji dokonały rewolucji nie tylko w komunikacji między uczonymi i w wyszukiwaniu informacji, ale także w podejściu do odkryć naukowych, tworząc wirtualne laboratoria i infrastrukturę dla usług obliczeniowych „na żądanie" w trybie przetwarzania sieciowego (ang. *grid*) lub w chmurze (ang. *cloud*).

Zamykając ten fragment wykładu, warto przytoczyć jeszcze jeden konkretny przykład kompleksowego działania sztucznej inteligencji dla wspomagania decyzji. Chodzi o wspomaganie osób niewidomych. Chieko Asakawa jest niewidomą Japonką-informatykiem, pracującą w IBM Research – Tokyo. Miałem okazję wysłuchać jej prezentacji w listopadzie 2019 r. podczas Światowego Forum Nauki w Budapeszcie. W aplikacji, którą przedstawia film[15], widzimy 6 typów zastosowań SI: orientacja przestrzenna i nawigacja – rozpoznawanie twarzy – rozpoznawanie emocji – rozpoznawanie kontekstu zdarzeń – rozpoznawanie tekstu – synteza mowy.

4. Generatywna sztuczna inteligencja

Nowym etapem rozwoju sztucznej inteligencji jest generatywna sztuczna inteligencja oparta na uogólnionych modelach zdolnych do wykonywania różnych zadań intelektualnych człowieka. Jest ona nadzieją dla nowoczesnych systemów zarządzania i wytwarzania, tzw. Przemysłu 4.0 (ang. *Industry 4.0 + AI = AI4Industry*). Charakteryzują się one strukturą modularną połączoną w jeden system zwany łańcuchem bloków (ang. *blockchain*). Łańcuch bloków to zdecentralizowana, rozproszona baza danych (transakcyjnych) bez centralnych komputerów, do której dostęp może uzyskać każdy uczestnik procesu. W bazie tej dokonywane jest gromadzenie i kompleksowa

[15] https://youtu.be/f-mQIWnO3Ag? si=EWxzCumV5lyRj59U

analiza danych z wielu źródeł. Taka struktura zastępuje klasyczną strukturę hierarchiczną z centralnym zarządzaniem i deterministycznym nadawaniem reguł komunikacji między składowymi systemu. Decentralizacja i autonomia podsystemów mają wzmóc odporność systemu na zmianę kontekstu i warunków operacyjnych. Powiązania między podsystemami tworzą się tak jak w sieci społecznościowej i rolą algorytmów sztucznej inteligencji jest odkryć je dla zidentyfikowania skupień wynikających z interakcji niekoniecznie z góry zadekretowanych (klastrów). Jest to typowe dla podejścia behawioralnego charakterystycznego dla sztucznej inteligencji: „pokaż mi przykłady działania, a algorytmy sztucznej inteligencji stworzą jego model".

Moduł systemu zarządzania lub wytwarzania może być traktowany jak autonomiczny agent zdobywający wiedzę o (wirtualnym) otoczeniu dzięki uczeniu się z danych. Agentem może być robot „ucieleśniający" sztuczną inteligencję lub program komputerowy działający według algorytmu uczącego. W tradycyjnym uczeniu maszynowym algorytmy uczenia się są typowo nastawione na pojedynczy dobrze zdefiniowany problem i nie dostosowują się do zmiennych okoliczności. Jest to uczenie się z funkcją nagrody i postępowanie w trybie 'sytuacja-akcja' według wyuczonych reguł maksymalizujących przyjętą funkcję użyteczności (ang. *extrinsic motivation*).

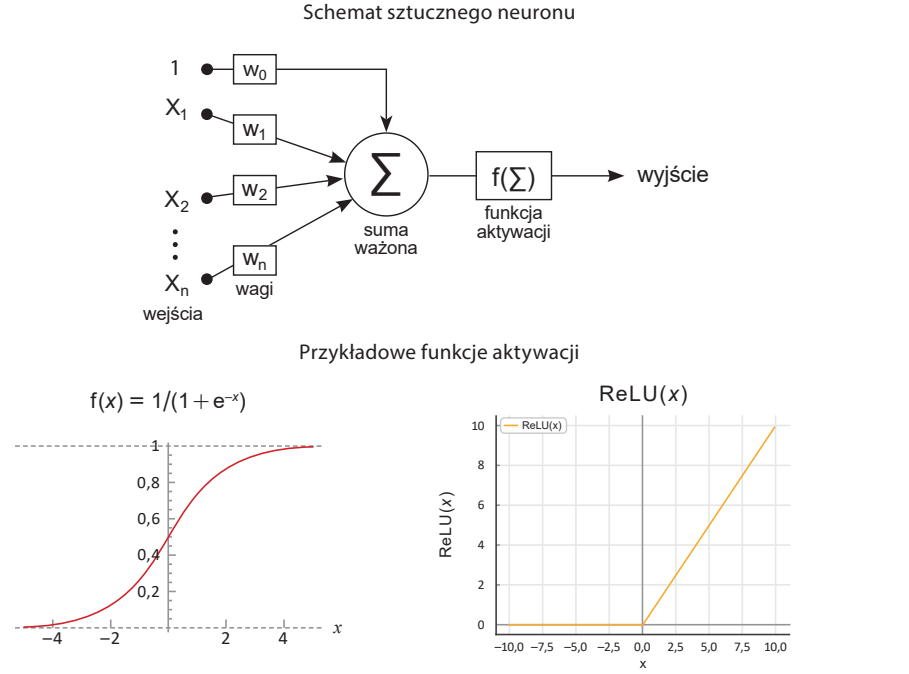
Nowy paradygmat uczenia maszynowego, który bardziej przystaje do systemu modularnego z niezadekretowanymi interakcjami między podsystemami, nazywa się uczeniem ze wzmocnieniem (ang. *reinforcement learning*). Uczenie ze wzmocnieniem odbywa się bez funkcji nagrody, a eksploracja otoczenia dokonuje się przez akcję z wyciąganiem wniosków z tego, co się stało (ang. *intrinsic motivation*). Jest to analog do programowania dynamicznego, gdzie przyszłe stany są liczne i nie wiadomo *a priori*, który jest najlepszy. Agent uczy się nie tyle strategii właściwego postępowania, ile sposobu oceny swojego działania i własnych preferencji. Ponieważ agent uczy się preferencji w kontekście wielu konfliktowych celów, wykorzystuje podejście wielokryterialne[16]. Taka aktywność sprzyja nabywaniu szerokich kompetencji zamiast specjalizacji w osiąganiu narzuconego celu. Te kompetencje tworzą elementy wiedzy, z których agent może zbudować rozwiązanie nowego problemu, jaki może się pojawić.

Inteligentny agent uczący się ze wzmocnieniem ma zatem przewagę nad agentem uczącym się według algorytmu klasycznego, gdyż jego wiedza pozwala na działanie dostosowane do wcześniej nieznanych scenariuszy, co jest bliższe racjonalnemu działaniu człowieka.

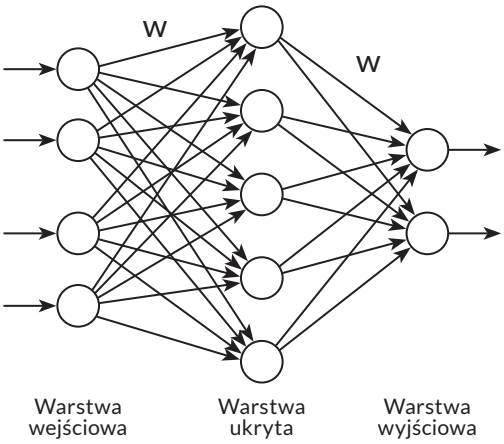
[16] E. Hüllermeier, R. Słowiński, Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies. *4OR – A Quarterly Journal of Operations Research*, 22 (2024) part I: 179–209, part II: 313–349.

Techniki uczenia maszynowego są ciągle udoskonalane, w szczególności uczenie złożonych funkcji matematycznych zwanych Sztucznymi Sieciami Neuronowymi (SSN) (ang. *Artificial Neural Network – ANN*). Tak zwane uczenie głębokie (ang. *deep learning*) jest kamieniem milowym nowoczesnej SI. Techniki głębokiego uczenia są bardzo efektywne w nauce wielowarstwowych sieci neuronowych do przetwarzania i rozumienia złożonych wzorców oraz relacji w danych obrazowych i językowych. Trenowanie modeli uczenia głębokiego, polegające na konfigurowaniu parametrów neuronów sieci, wymaga intensywnych obliczeń, które można łatwo zrównoleglić. Z tego powodu stosuje się procesory graficzne (GPU), które dzięki swojej równoległej strukturze są niezwykle efektywne w przyspieszaniu uczenia głębokiego. Duże Modele Językowe (DMJ) (ang. *Large Language Models – LLM*), takie jak ChatGPT-4, Claude, PaLM, Liama, LaMDA, Gemini, Cohere, Orca itd., opierają się na technikach uczenia głębokiego. Tworzą one klasę tzw. generatywnych modeli SI.

Przypomnijmy, że sztuczny neuron to wieloparametryczna funkcja matematyczna, przedstawiona na rys. 1. Sieci tworzą neurony podobnego typu, tak jak



Rys. 1. Sztuczny neuron – wagi w są parametrami połączeń (synaps) między neuronami



Rys. 2. Perceptron wielowarstwowy

to przedstawiono na rys. 2. Perceptron wielowarstwowy może być wykorzystywany do klasyfikowania zbiorów obiektów, które nie są liniowo separowalne. Jego pomysłodawcą jest Geoffrey Hinton, drugi obok Johna Hopfielda laureat Nagrody Nobla z fizyki w 2024 r. Hopfield wprowadził tzw. asocjacyjne sieci neuronowe dające możliwość rekonstrukcji i rozpoznawania wcześniej zapamiętanych wzorców na podstawie skojarzeń, które bazują na dostępnym fragmencie wzorca. Idea trenowania, czyli uczenia SSN, polega na wyznaczaniu wartości wag (parametrów SSN), które gwarantują poprawne odtwarzanie relacji wejście-wyjście zachodzących w przykładach uczących. Do wyznaczania tych parametrów służą algorytmy optymalizacji rozwijane od dziesięcioleci w badaniach operacyjnych (BO), co podkreśla ścisły związek BO i SI.

Duże modele językowe są trenowane na ogromnych i zróżnicowanych zbiorach danych tekstowych, co pozwala im generować spójny i kontekstowo odpowiedni tekst na podstawie otrzymanych danych wejściowych. Potrafią też śledzić kontekst rozmowy.

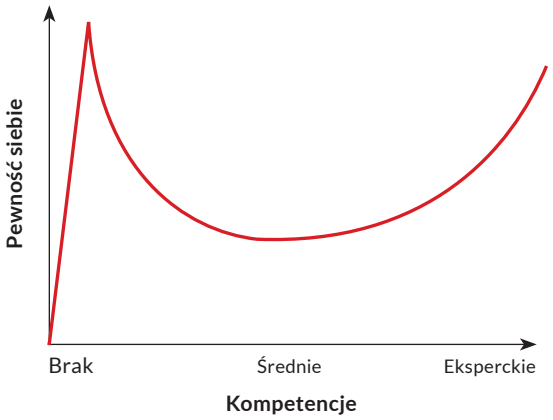
Najsłynniejszy dziś DMJ to ChatGPT firmy OpenAI. Wyjaśniając jego nazwę, *Generative Pre-trained Transformer*, to sieć neuronowa zaprojektowana do nauki odległych zależności statystycznych w tekstach. Trenowanie modelu językowego obejmuje techniki przewidywania następnego tokenu (ang. *Next-Token-Prediction – NTP*) oraz maskowania języka (ang. *Masked-Language-Modeling – MLM*). W technice MLM pewne słowa w sekwencji są „maskowane", a model jest trenowany, aby przewidzieć poprawne słowa na podstawie kontekstu otaczającego

tekstu. Na przykład, dla wyrażenia „Kot siedział na...", model szacuje prawdopodobieństwo kolejnego słowa, takiego jak „kanapie", „podłodze" lub „dachu", biorąc pod uwagę częstotliwość występowania tych słów w innych tekstach.

Modele językowe nie korzystają z żadnego logicznego modelu wnioskowania o świecie. Potrafią zasugerować, jak odpowiedź powinna brzmieć, co różni się od tego, jaka odpowiedź powinna być.

Nawet twórcy ChatGPT z OpenAI przyznają, że ludzie mylą wydajność z kompetencją. DMJ nie są inteligentne w taki sposób, jak nasz mózg. Nie uczą się informacji w uporządkowany sposób. W przeciwieństwie do ludzi, przynajmniej na razie, nie mogą wchodzić w interakcje ze światem, aby uczyć się przyczynowości, kontekstu czy analogii. Co więcej, mówi się, że „halucynują", czyli tworzą stwierdzenia, które na pierwszy rzut oka wydają się wiarygodne, ale są bezpodstawne lub zafałszowane błędnymi danymi spotykanymi w sieci.

Do DJM można metaforycznie zastosować tzw. efekt Dunninga-Krugera, znany z psychologii, który przedstawia rys. 3. Tak jak ludzie o niskich kompetencjach, DMJ generują bardzo pewne odpowiedzi, nawet jeśli brakuje im wystarczających informacji lub gdy dane są niejasne bądź nieprecyzyjne. Ta „pewność siebie" wynika ze sposobu, w jaki trenowane są DMJ do tworzenia odpowiedzi w sposób płynny i autorytatywny. Podobnie jak u niekompetentnych ludzi, DMJ nie wiedzą, czego nie wiedzą. Efekt Dunninga-Krugera może odnosić się także do rozumienia DMJ przez użytkowników. Osoby nieświadome ograniczeń SI mogą zakładać, że DMJ „rozumieją" informacje tak jak ludzie i są wiarygodnym źródłem wiedzy. Takie przecenianie sprawia, że użytkownicy zbytnio polegają na wynikach genero-



Rys. 3. Efekt Dunninga-Krugera

wanych przez DMJ, nie poddając ich krytycznej ocenie. Podobnie jak ludzie zdobywający doświadczenie, DMJ mogą ulepszać swoje działanie dzięki dodatkowym danym i dostrajaniu. Wtedy ich wiarygodność wzrasta.

DMJ są trenowane nie tylko na tekstach, ale także obrazach, filmach, kodach komputerowych i dźwiękach. Stąd pojawiło się ogólniejsze pojęcie: Duże Modele Multimodalne (DMM) (ang. *Large Multimodal Models – LMM*). Dostępne są liczne modele specjalizowane, wywodzące się z DMJ, zwane wtyczkami (ang. *plug-ins*). Wtyczki te dokonują np. podsumowań tekstów z zadaną liczbą znaków, wyjaśniają i dokumentują kod programisty, tworzą obraz lub wideo na słowne polecenie, imitują mowę, tworzą muzykę, piosenki i wiersze oraz rozwiązują testy z fizyki i matematyki. ChatGPT-4o1 miał 83% sukcesu przy rozwiązywaniu zadań z olimpiady matematycznej. Dziś ocenia się jego poziom kompetencji na porównywalny z poziomem doktoranta.

Powstał też polski DJM Bielik, ale jest niedoinwestowany, a także polski serwis porad prawnych Legitio. Wtyczka Scholar AI do ChatGPT-4o podaje informacje o publikacjach naukowych, ich autorach oraz o cytowaniach.

DMJ lepiej niż ludzie rozpoznają ze zdjęć osób ich preferencje seksualne i polityczne. System SORA od OpenAI tworzy bardzo realistyczne klipy wideo na polecenia słowne, np. utwórz klip „spacer po Tokio". DMM są już instalowane na smartfonach, jak Apple Intelligence, Galaxy AI, do generowania obrazów, modyfikacji zdjęć, rozpoznawania twarzy, tłumaczenia i podsumowywania notatek lub nagrań mowy.

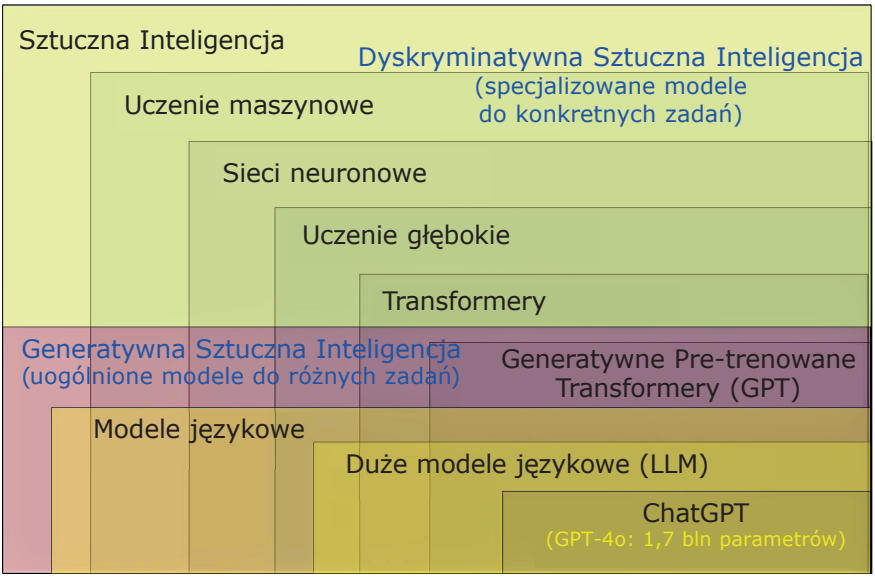
Generatywna SI ma też ujemne strony. Deepfake Porn[17], na podstawie jednego wyraźnego zdjęcia twarzy, jest w stanie w ciągu mniej niż 25 min stworzyć za darmo 60-sekundowe pornograficzne wideo.

Liderami w rozwoju DMJ i DMM są OpenAI, Google i Meta. Sporo ich produktów jest otwartych, co ułatwia deweloperom tworzenie niezliczonej liczby modyfikacji i specjalizacji. Pojawiła się także nowa klasa modeli SI – DMM wieloagentowe, w których wiele modeli współpracuje dla rozwiązania złożonych zadań.

Schemat podany na rys. 4 podsumowuje użyte dotąd pojęcia związane ze SI i przedstawia ich hierarchię (rozmiar okien nieistotny).

Warto zwrócić uwagę, jak duże są koszty trenowania i używania DJM. Modele uczenia głębokiego to „brutalne" techniki statystyczne, które uczą się na ogromnych ilościach danych (rzędu 10 TB tekstu). Sieci neuronowe DMJ zawierają setki warstw i biliony (10^{12}) wag, co sprawia, że obliczenia są niezwykle energochłonne.

[17] E. Strickland, Deepfake Porn is Leading to a New Protection Industry, *IEEE Spectrum*, 15 July 2024.



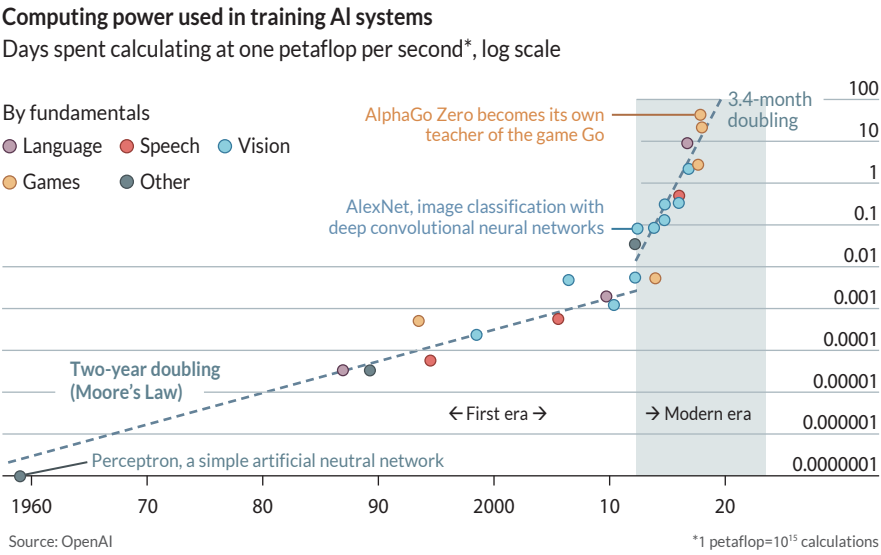
Rys. 4. Podsumowanie i hierarchia pojęć związanych ze sztuczną inteligencją

ChatGPT ma 1,7 bln parametrów uczenia maszynowego. Trenowanie modelu GPT-4 wymaga użycia 25 tys. jednostek GPU przez 90–100 dni, przy kosztach wynoszących 100 mln dolarów, a zużycie energii wynosi od 51 773 do 62 319 MWh (energia zużywana przez 1000 gospodarstw domowych w USA przez 5–6 lat). Zużycie energii przy użytkowaniu jest 10 razy wyższe niż koszty treningu (w ciągu pierwszego miesiąca 2024 r. ChatGPT zużył tyle prądu, co 26 tys. domostw)[18].

Wysokie koszty energii sprawiają, że DMJ i DMM są rozwijane przez bogate firmy, a nie przez laboratoria uniwersyteckie. Na przykład właściciel Facebooka, firma META, inwestuje około 90 mld dolarów rocznie, w tym 10 mld dolarów w rozwój Metaverse, w nadziei, że wirtualna i rozszerzona rzeczywistość (ang. *Virtual Reality – VR*, *Augmented Reality – AR*) staną się kolejną dużą platformą obliczeniową. Metaverse ma być następnym krokiem milowym na drodze rozwoju Internetu, tworząc wirtualny lub hybrydowy świat trójwymiarowy wygenerowany cyfrowo, w którym społeczność będzie mogła odnajdywać się dla pracy lub rozrywki za pomocą gogli VR, kontrolerów ruchu i dźwięku przestrzennego.

[18] arXiv:1906.02243.

Od 2012 r. moc obliczeniowa potrzebna do trenowania systemów SI podwaja się co 3–4 miesiące (rys. 5).



Rys. 5. Moc obliczeniowa wykorzystywana do trenowania systemów SI[19]

Z powyższych powodów powstaje nowa generacja komputerów na potrzeby trenowania DMJ: to komputery neuromorficzne, które w przeciwieństwie do konwencjonalnych komputerów, używających oddzielnych procesorów do obliczeń i pamięci do przechowywania danych, integrują te dwie funkcje w sztucznych neuronach. Dzięki temu nie ma potrzeby kosztownej synchronizacji. Ten sposób przetwarzania informacji jest bliższy naszym mózgom. Na przykład, komputer Hala Point firmy Intel przetwarza informacje 50 razy szybciej, zużywając 100 razy mniej energii niż zwykłe komputery. Mówi się o 15 bln operacji na sekundę i na wat. Hala Point jest wyposażony w 1152 nowe procesory Intela, a komputer może obsłużyć 1,15 mld sztucznych neuronów i 128 mld synaps, co odpowiadałoby mózgowi małego ssaka. Mózg ludzki ma około 100 mld neuronów i 100 bln połączeń-synaps. Nasz mózg ma ponadto wyspecjalizowane regiony i kontrolę emocji.

[19] https://openai.com/research/ai-and-compute

5. Inteligencja maszyn jako racjonalne działanie motywowane planem i wiedzą o otoczeniu

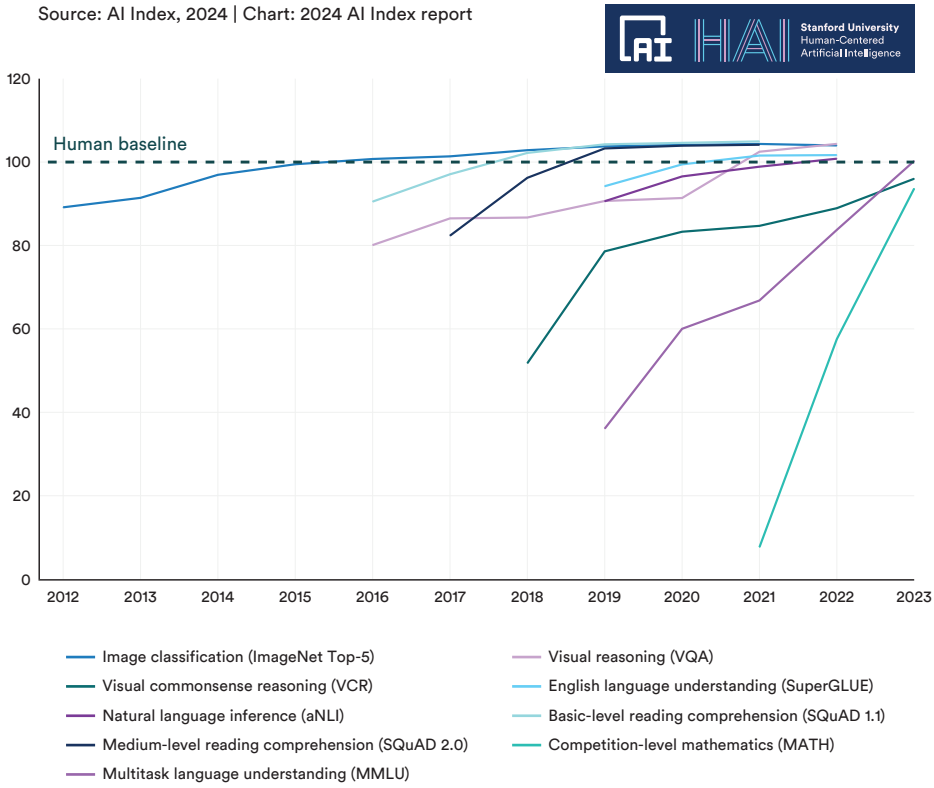
Przechodząc do trzeciej definicji inteligencji maszynowej, która ma ambicje dorównać ludzkiemu rozsądkowi przez autonomiczne zdobywanie wiedzy o otoczeniu i działanie planowe, musimy zadać pytania o świadomość maszyn i możliwość ich „uczłowieczenia" przez nabycie cech i uczuć wyższych, którymi człowiek odróżnia się na przykład od zwierząt. Jakie konsekwencje dla człowieczeństwa niesie rozwój SI?

Najprostsze (choć bezwzględnie bardzo trudne) pytanie tego typu to, czy maszyna może osiągnąć stan samoświadomości, która pozwoli jej zdać sobie sprawę z relacji „ja-środowisko" i powiedzieć: „ja jestem". Stanisław Krajewski[20] zauważa, że uświadomione „ja" w relacji do przedmiotów jest łatwiejsze do uświadomienia przez maszynę niż „ja" w relacji do innych ludzi, czyli „ja-ty". Te relacje są o wiele trudniejsze do zasymulowania niż relacje robota ze środowiskiem, które nie ma cech ludzkich. Relacje współpracy, życia razem, relacje uczuciowe, odczuwanie wstydu, nadziei czy wdzięczności są niezwykle trudne do zasymulowania. Ponadto wdzięczność czy nadzieja są pojęciami przekraczającymi rzeczywistość „tu i teraz", a zatem posiadają element transcendencji charakterystyczny dla religii i modlitwy. Rodzą one szczególny rodzaj relacji „ja-Bóg". Można wątpić, czy maszyny za sprawą programów komputerowych wyewoluują do takiego poziomu świadomości.

Niemniej jednak robot wykorzystujący generatywną sztuczną inteligencję potrafi malować, komponować muzykę czy głosić kazania, jak „Mindar" do buddystów. SI dorównuje już, a nawet przewyższa niektóre zdolności rozumowe ludzkiego mózgu. Raport „2024 AI Index Report" przedstawiony przez Human-Centered Artificial Intelligence Unit Uniwersytetu Stanforda (rys. 6) pokazuje, że SI przewyższyła ludzkie możliwości w kilku benchmarkach, w tym w klasyfikacji obrazów, rozumowaniu wizualnym i rozumieniu języka angielskiego. Jednak wciąż pozostaje w tyle w przypadku bardziej złożonych zadań, takich jak matematyka na poziomie konkursowym, wizualne rozumowanie zdroworozsądkowe i planowanie.

Moim zdaniem możliwe jest, że maszyny rozwiną swoją własną inteligencję, która pozwoli im zachowywać się w taki sposób, że ludzie subiektywnie oceniający ich świat będą mogli powiedzieć „one zachowują się, jakby były świadome/

[20] St. Krajewski, Can a robot be grateful? Beyond Logic, towards religion, *Eidos – Journal for Philosophy of Culture*, 4, 6 (2018) 4–13.



Rys. 6. Raport z testów porównania SI w konkurencji z ludzkim mózgiem[21]

wdzięczne/kochające". Taka relacja i ocena będą podobne do tego, co obecnie mówimy o zwierzętach domowych. Dyskusja „czy one naprawdę mają świadomość" i „czym różnią się od ludzi" będzie analogiczna do tej o zwierzętach.

Sądzę także, że maszyny przejmą ludzką świadomość jedynie w sensie behawioralnym – nauczą się naszych upodobań i reakcji. Jeśli maszyny stworzą swój własny świat „myśli", to nie będziemy mieli do niego dostępu, tak jak nie mamy dostępu do „myśli" naszego psa – to po prostu będzie inna inteligencja. W poprzednim rozdziale podkreślaliśmy, że w sztucznej inteligencji nowej generacji autonomiczny

[21] https://aiindex.stanford.edu/report/

agent uczy się przez eksplorację środowiska z wyciąganiem własnych wniosków ze skutków swoich akcji i nazywaliśmy to uczeniem ze wzmocnieniem. Doświadczenie to kształtuje swoistą świadomość agenta, której my możemy nie zrozumieć. Oceniamy go po czynach. Stąd moja konkluzja, że maszyny jako sztuczni agenci nie zostaną „uczłowieczeni", ale prawdopodobnie wytworzą porozumienie między sobą, które pozwoli im na autonomiczne działanie zespołowe.

Powyższy pogląd rezonuje z tym, co przytoczyliśmy już na pierwszych stronach tego artykułu za Stanisławem Krajewskim[22]: „Twierdzenie Gödla jednak dopuszcza możliwość, że taki program lub robot [dorównujący lub przewyższający ludzki umysł – przyp. RS] mógłby powstać w wyniku działania jakiejś superinteligencji lub w wyniku ewolucji w społeczeństwie robotów. W takim przypadku nie bylibyśmy jednak w stanie odkryć ani potwierdzić tej równoważności".

Czy zatem powinniśmy się bać, że stworzymy byt, który nas intelektualnie przerośnie i ubezwłasnowolni? Stephen Hawking pod koniec życia przestrzegał przed badaniami w kierunku stworzenia samodzielnej sztucznej inteligencji. Według niego moment, w którym uzyskamy ostateczną formę sztucznej inteligencji – samodzielną i samoświadomą będzie najgorszy w historii ludzkości.

Futuryści w rodzaju Nicka Bostroma[23] zapowiadają zbliżanie się tzw. punktu osobliwości technologicznej (ang. *singularity*), w którym algorytmy samodzielnie podejmą zadanie świadomego samorozwoju i określą jego kierunki. Ma to doprowadzić do powstania silnej, ogólnej sztucznej inteligencji, która stworzy samoprojektujące się maszyny.

Powyższy pogląd wspierany jest także mirażem zmapowania ludzkiego mózgu w pamięć maszyny. Elon Musk, właściciel firmy produkującej samochody elektryczne Tesla, finansuje startup „Neuralink" opracowujący implant do mózgu, który ma pozwolić na przeniesienie wspomnień z mózgu do chmury, czyli – mówiąc w uproszczeniu – przeniesienie naszej świadomości do robota. Wtedy zdolność efektywnego obliczania, radzenia sobie z ogromną liczbą danych, w połączeniu z ludzką inteligencją miałaby stworzyć superinteligencję. Ponadto wydobycie wspomnień z biologicznego mózgu oznaczałoby życie wieczne.

Faktycznie, uwiedzeni tą ideą prorocy sztucznej inteligencji postulują nowe niby-religie: „Bóg został uśmiercony i człowiek stanął na jego miejscu"[24]. To brzmi równie złowieszczo jak pokusa z Księgi Rodzaju: „będziecie jak Bóg".

---

[22] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

[23] N. Bostrom, *Superinteligencja: scenariusze*, *strategie*, *zagrożenia*, Helion, 2016.

[24] J. Koronacki, Cyberprzestrzeń, sztuczna inteligencja i posthumanizm, *Teologia Polityczna*, 2019-09-30.

Yuval Harari, autor książki pt. *Homo deus*, propaguje koncepcję dataizmu, która jest systemem wierzeń skoncentrowanym wokół siły i wartości danych. W tym paradygmacie dane postrzegane są jako najważniejszy zasób, a ich przetwarzanie i przepływ uważa się za cechę definiującą inteligencję i kluczowy czynnik w podejmowaniu decyzji. Dataizm zakłada, że wszechświat składa się z przepływów danych, a wartość każdego zjawiska lub bytu jest określana przez jego wkład w przetwarzanie danych. Dataizm przekazuje kompetencje podejmowania decyzji z ludzi na algorytmy, ponieważ one lepiej interpretują złożone wzorce odkryte z danych. To wyraźne zagrożenie dla ludzkiej wolnej woli.

Musimy zdać sobie sprawę, że SI stanowi wyzwanie dla głębokich dyskusji filozoficznych i religijnych dotyczących istoty człowieczeństwa. Dużo trafnych pytań nt. przyszłości SI stawia Papież Franciszek w niedawnym Orędziu na Dzień Środków Przekazu: „Jak zapewnić przejrzystość procesów informacyjnych? […] Jak zapobiec sprowadzaniu wielu źródeł tylko do jednego źródła, do algorytmicznie opracowanego jednolitego myślenia? I jak, z drugiej strony, możemy wspierać środowisko, które zachowuje pluralizm i reprezentuje złożoność rzeczywistości? […] czy sztuczna inteligencja doprowadzi do utworzenia nowych kast opartych na dominacji informacyjnej, rodząc nowe formy wyzysku i nierówności; czy wręcz przeciwnie, przyniesie więcej równości, promując poprawną informację i większą świadomość przeżywanych obecnie przemian dziejowych […]. Z jednej strony pojawia się widmo nowego niewolnictwa, z drugiej zdobycie wolności; z jednej strony możliwość warunkowania myślenia wszystkich przez nielicznych, a z drugiej możliwość uczestniczenia wszystkich w procesie myślowym” [25].

W tym kontekście przychodzi na myśl profetyczna powieść Aldousa Huxleya pt. *Nowy wspaniały świat* z 1931 r., która ukazuje społeczeństwo przyszłości, w którym nauka, technologia i biotechnologia zostały wykorzystane do stworzenia idealnego, ale jednocześnie odczłowieczonego społeczeństwa. W tym świecie jednostki są kontrolowane już od najwcześniejszych etapów życia, od ich genetycznego projektowania (eugenika) po indoktrynację ideologiczną.

Na tej fali powstał pretensjonalny ruch zwany transhumanizmem[26]. Atrakcyjność transhumanizmu wynika z faktu, że oferuje on nadzieję na spełnienie wiecznego marzenia o raju na ziemi poprzez przekształcenie człowieka i świata za pomocą narzędzi naukowych i technologicznych. Transhumanizm obiecuje trzy razy super: superdługowieczność, superszczęście i superinteligencję. Manifest

[25] https://www.vatican.va/content/francesco/pl/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html

[26] Transhumanizm, *ETHOS*, 28, 3 (2015).

transhumanistów przypomina manifest Kominternu[27]: „Transhumaniści świata, łączcie się – mamy do zdobycia nieśmiertelność, a do stracenia jedynie biologię. Razem możemy przełamać łańcuchy biologii i przekroczyć ograniczenia związane z niedoborem, płcią, wiekiem, etnicznością, rasą, śmiercią, a być może nawet czasem i przestrzenią. […] By zapewnić większe szanse przeżycia, uzyskać kontrolę nad naszym własnym przeznaczeniem i być wolnym, musimy zapanować nad ewolucją".

Biolog Julian Huxley (brat Aldousa, choć jego ideowe przeciwieństwo) pisał już w 1957 r.: „Rodzaj ludzki, o ile tego zapragnie, może dokonać transcendencji siebie"[28]. Transhumanizm w swej radykalnej formie okazuje się projektem zmierzającym do modyfikacji człowieka i przebudowy społeczeństwa. Ciekawą analizę tego zjawiska przedstawił Jacek Koronacki[29].

Media chętnie podsycają wyobraźnię w kierunku transhumanizmu. Na przykład, w 2014 r. Johnny Depp wcielił się w filmie pt. *Transcendence* w niezwykłą rolę człowieka, którego umysł zostaje przeniesiony do maszyny. Pojawia się jako holograficzny obraz człowieka „żyjącego w komputerze".

Dodajmy jednak, że pomimo przedstawionego powyżej postępu SI, realizacja różnych entuzjastycznych obietnic nie jest tak szybka, jak niedawno przewidywano. Na przykład, powszechność pojazdów autonomicznych została zapowiedziana już 9 lat temu. Tymczasem autonomiczne taksówki wciąż są w fazie testów i nie przynoszą zysków firmom, które je wprowadzają.

Jeszcze jedna ważna uwaga. Mówi się, że demokracja jest za sprawą sztucznej inteligencji zagrożona algokracją – algorytmy mają o nas decydować, jako niekorumpowalne, obiektywne i przestrzegające reguły prawa. W niektórych krajach obserwujemy kontrowersyjne praktyki polegające na monitorowaniu i ocenianiu obywateli w różnych sferach życia, co może naruszać prywatność i swobody obywatelskie. Zachowanie ludzi oceniane jest według pewnych procedur skoringowych, zgodnie z którymi obywatel otrzymuje na starcie określoną liczbę punktów i za dobre sprawowanie zyskuje punkty, a za występki, takie jak zapalenie papierosa w miejscu niedozwolonym, traci punkty. System skoringowy pochodzi z systemów bankowych – kiedy klient udaje się po kredyt, bank zbiera różne informacje o jego dochodach, przeszłości kredytowej, stylu życia, a algorytm podaje wynik w skali od zera do stu, który decyduje, czy klient kwalifikuje się na dany kredyt. Taki system przechodzi powoli do oceny globalnej obywateli, zwłaszcza

[27] Transhumanism manifesto, https://www.singularityweblog.com/a-transhumanist-manifesto/

[28] M. Grabowski, Transhumanizm: geneza-założenia-krytyka, *ETHOS*, 28, 3 (2015) 23–41.

[29] J. Koronacki, O końcu historii i transhumanizmie, *Teologia Polityczna*, 2017–01–29.

w krajach niedemokratycznych. Czytamy, że systemy oceny ludzi zaczęły działać w Chinach[30] i w Indiach[31]. Powstają wielkie bazy danych zawierające informacje o wszystkich obywatelach, razem z ich wizerunkami, odciskami palców itd. Dzięki bardzo wydajnym algorytmom rozpoznawania twarzy oraz ogromnej liczbie kamer na ulicach i w obiektach możliwe jest znalezienie osoby nawet w wielkim tłumie. Chociaż system ten został wprowadzony z uzasadnieniem przydatności do walki z korupcją lub terroryzmem, jego negatywne skutki dla społeczeństwa są łatwe do przewidzenia.

Boimy się braku transparentności algorytmów i dlatego coraz głośniej mówi się o wyjaśniających modelach sztucznej inteligencji. To jasne, że sposób działania algorytmów sztucznej inteligencji może opierać się na innych zasadach niż ludzkie procesy myślowe. W naszej pracy z histopatologami spotkaliśmy się np. z cechą preparatu nazwaną „skórą tygrysią". Jak ją wytłumaczyć komputerowi? Skończyło się na przyjęciu cechy długości fraktalnej poziomic gęstości jąder komórkowych[32], która to cecha nie jest zrozumiała dla człowieka. W innym naszym projekcie dotyczącym wspomagania diagnostyki w izbie przyjęć szpitala pediatrycznego zalecenia musiały mieć czytelną postać reguł logicznych, a nie wyniku numerycznego, czyli tzw., medycznego skoringu w skali numerycznej[33].

Przed zakończeniem wspomnę jeszcze o pewnym nietypowym zastosowaniu sztucznej sieci neuronowej w uczeniu głębokim. Niedawno ukazało się doniesienie o ciekawym zastosowaniu głębokiego uczenia splotowej sieci neuronowej, zwanej CNN[34], do sprawdzania autentyczności obrazów znanych mistrzów. Analizowano obraz *Salvator Mundi*, przypisywany Leonardo da Vinci (sprzedany na aukcji Christie's w 2017 r. za 450 mln dol.), oraz dwa obrazy Rembrandta, *Saskia* i *Mężczyzna w złotym hełmie*. Sieć CNN, ucząc się stylu malowania Leonarda da Vinci i Rembrandta na innych niepodważalnych obrazach tych mistrzów, stwierdziła, że fragmenty twarzy Saskii, żony Rembrandta, były prawdopodobnie przemalowane po Rembrandcie przez kogoś innego. Z kolei sieć CNN silnie zaklasyfikowała *Mężczyznę w złotym hełmie* jako obraz Rembrandta, pomimo że ekspert holenderski w 1985 r. zawyrokował, że jest to obraz z kręgu Rembrandta (znajduje się on

[30] https://www.computerworld.pl/news/Permanentna-inwigilacja-czyli-scoring-po-chinsku, 411813.html

[31] https://www.rp.pl/Plus-Minus/303069985-Indie-chca-inwigilowac-tak-jak-Chiny.html

[32] J. Jelonek, K. Krawiec, R. Słowiński, J. Szymaś, *Grizzly* – an image analysis and classification system oriented towards medical applications. *Journal of Decision Systems*, 7 (1998) 39–53.

[33] W. Michalowski, R. Słowiński, Sz. Wilk: MET System – a New Approach to m-Health in Emergency Triage. *The Journal on Information Technology in Healthcare*, 2 (2004) 237–249.

[34] S.J. Frank, This convolutional neural network can tell you whether a painting is a fake. *IEEE Spectrum*, 29 September 2021.

w Galerii Gemäldegalerie w Berlinie). U *Salvatora Mundi* wątpliwości wzbudzało tło i ręka błogosławiąca, co potwierdza fakt, że obraz był poważnie uszkodzony, a jego części zrekonstruowane.

Sieć neuronowa ma jednak tę wadę, że nie wyjaśnia swoich decyzji. W warstwach pośrednich tej sieci tworzą się metacechy, które nie są dla człowieka zrozumiałe – ten efektywny klasyfikator pozostaje dla nas „czarną skrzynką". Znacznie bardziej transparentne są inteligentne algorytmy wspomagania decyzji oparte na modelach w postaci reguł logicznych[35], lecz za czytelność rekomendacji płacimy tu efektywnością uczenia się na dużych zbiorach danych.

Dotykamy tutaj poważnego tematu odpowiedzialności za decyzje algorytmów sztucznej inteligencji. Na przykład, gdy użytkowany przez nas pojazd autonomiczny znajdzie się w sytuacji kryzysowej, w której będzie musiał dokonać wyboru mniejszego zła, to kto odpowie za powstałe konsekwencje – my, jako właściciel pojazdu, czy projektant autonomicznego systemu sterowania? Musimy zatem wiedzieć, jaki system wartości został zakodowany w algorytmie sterującym. Musimy kontrolować to, co maszyny nam zalecają. W związku z tym wszystkie deklaracje etyczne podkreślają dzisiaj istotność wyjaśnialności modeli sztucznej inteligencji. Niedawno Komisja Europejska opublikowała raport ekspertów w sprawie „Wytycznych w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji"[36]. Raport wskazuje na moralną i prawną odpowiedzialność jej twórców i apeluje o tworzenie transparentnych systemów, dających wgląd w rekomendacje i ich motywacje.

Powyższe rozważania prowadzą do naturalnego pytania, czy postęp technologiczny w zakresie sztucznej inteligencji, zamiast wspomagać człowieka, może go ubezwłasnowolnić? Moim zdaniem tak, jeśli bezkrytycznie i bez zrozumienia podpowiedzi maszyny polegalibyśmy na jej decyzjach – wtedy SI ograniczy ludzką wolną wolę. Póki co, wszystko zależy od samego człowieka, bo maszyny i roboty nie przejęły (jeszcze) nad nami kontroli. W postępie technologicznym mają udział oba oblicza natury człowieka – dobre i złe – stąd postęp stanowi wypadkową walki tych sił w każdym z nas. Wobec tego najważniejszy jest wzrost samego człowieka, nie tylko w zakresie jego wiedzy, ale także ducha, który wspomaga rozpoznanie dobra i zła. Innymi słowy, możemy nie obawiać się sztucznej inteligencji, jeśli nie zaniedbamy rozwoju naszej własnej inteligencji i prawego sumienia.

---

[35] J. Błaszczyński, S. Greco, R. Słowiński, Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence*, 25, 2 (2012) 284–294.

[36] https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11–06/Ethics-guidelines-AI_PL.pdf