Nick Bostrom

# Superinteligencja

*Scenariusze, strategie, zagrożenia*

Bestseller „New York Timesa"

onepress
Helion

Tytuł oryginalny: Superintelligence: Paths, Dangers, Strategies, First Edition

Tłumaczenie: Dorota Konowrocka-Sawa

Projekt okładki: ULABUKA
Illustracja na okładce: ULABUKA

ISBN: 978-83-289-0327-2



Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
*https://helion.pl/user/opinie/supivv*
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

# Spis treści

ROZDZIAŁ 1.

# Dotychczasowe dokonania i obecne możliwości

**Zaczniemy od spojrzenia wstecz. Gdy przyjmie się najszerszą możliwą perspektywę, historia wydaje się ujawniać sekwencję wyraźnie wydzielonych faz wzrostu, przy czym każda z nich jest gwałtowniejsza od poprzedniej. Przywołujemy ten wzorzec, byśmy mogli sobie uzmysłowić, że możliwa jest kolejna (jeszcze szybsza) faza wzrostu, lecz nie przywiązujemy nadmiernej wagi do tej obserwacji, gdyż nie jest to książka na temat „przyspieszenia technologicznego", „wzrostu wykładniczego" ani żadnego z rozmaitych pojęć wrzucanych czasem do wspólnej kategorii „osobliwości" (ang. *singularity*). Następnie przyjrzymy się dotychczasowej historii rozwoju sztucznej inteligencji, po czym zbadamy współczesne możliwości uzyskane w tej dziedzinie. Na koniec rzucimy okiem na kilka najnowszych sondaży opinii ekspertów i zadumamy się nad naszą niewiedzą dotyczącą umiejscowienia przyszłych postępów w czasie.**

## Fazy wzrostu w dziejach ludzkości

Jeszcze kilka milionów lat temu nasi przodkowie zwisali sobie na gałęziach drzew pod sklepieniem afrykańskiej dżungli. W geologicznej czy nawet ewolucyjnej skali czasu do powstania gatunku Homo sapiens z ostatniego wspólnego przodka ludzi i małp człekokształtnych doszło bardzo niedawno. Przyjęliśmy postawę pionową, wykształciliśmy przeciwstawne kciuki i — co najistotniejsze — staliśmy się beneficjentami pewnych stosunkowo drobnych zmian w wielkości mózgu i organizacji neuronów,

które doprowadziły do ogromnego skoku zdolności poznawczych. W konsekwencji istoty ludzkie zyskały zdolność myślenia abstrakcyjnego, przekazywania złożonych idei i akumulowania informacji kulturowej z pokolenia na pokolenie w znacznie większym stopniu niż jakikolwiek inny gatunek żyjący na naszej planecie.

Te zdolności pozwoliły ludziom na rozwój coraz efektywniejszych technologii produkcji, umożliwiając naszym przodkom migracje i rozprzestrzenianie się na tereny odległe od lasu deszczowego i sawanny. Po rewolucji neolitycznej, kiedy to doszło do opracowania technik rolniczych, gęstość zaludnienia wzrastała wraz ze zwiększaniem się całkowitej liczby ludności Ziemi. Większa liczba ludzi równała się większej liczbie idei, zaś większa gęstość zaludnienia oznaczała, że idee te mogą się rozprzestrzeniać szybciej, a niektóre jednostki mogą się poświęcić rozwojowi wyspecjalizowanych umiejętności. Te zjawiska spowodowały podniesienie stopy wzrostu produktywności gospodarczej i potencjału technicznego. Kolejne wydarzenia, związane tym razem z rewolucją przemysłową, wywołały porównywalnie raptowną zmianę tempa tego wzrostu.

Takie zmiany tempa wzrostu mają istotne konsekwencje. Kilkaset tysięcy lat temu, w czasach prehistorycznych, rozwój był tak powolny, że potrzeba było rzędu miliona lat, by wzrost ludzkich zdolności produkcyjnych pozwolił na utrzymanie się przy życiu dodatkowego miliona ludzi. Pięć tysięcy lat przed naszą erą, po wybuchu rewolucji neolitycznej, stopa wzrostu podniosła się do takiego poziomu, że analogiczna zmiana wymagała już zaledwie dwustu lat. Dziś, po rewolucji przemysłowej, światowa gospodarka osiąga taki wzrost w zaledwie półtorej godziny[1].

Jeśli takie tempo wzrostu zostanie utrzymane przez stosunkowo długi okres, to nawet na obecnym poziomie przyniesie imponujące rezultaty. Jeśli światowa gospodarka będzie wzrastać w tym samym tempie, w jakim rosła przez ostatnie pięćdziesiąt lat, to do 2050 roku świat będzie mniej więcej 4,8 razy bogatszy niż dziś, a do 2100 roku mniej więcej 34 razy bogatszy niż dziś[2].

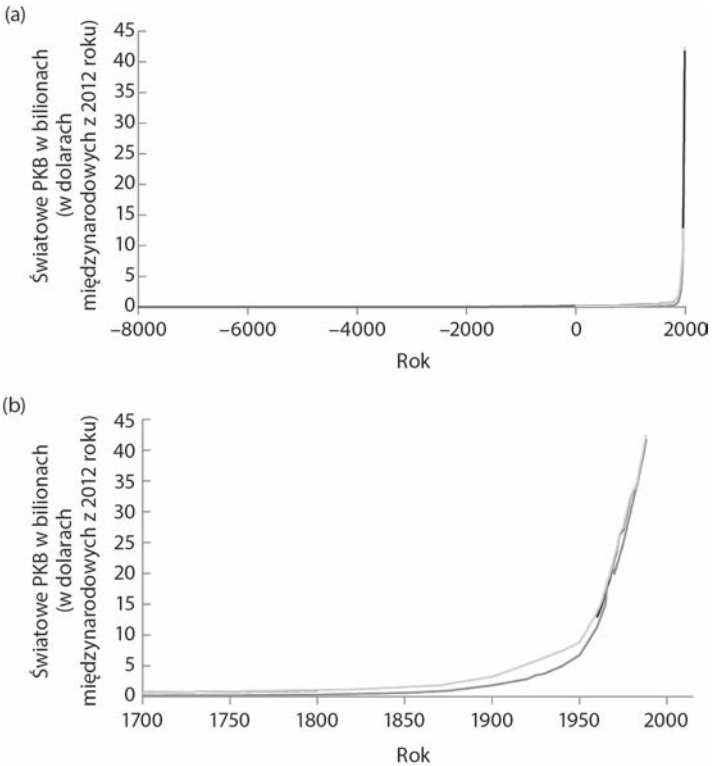
Jednak perspektywa stałego, nieprzerwanego wzrostu wykładniczego blednie w zestawieniu z tym, co może się zdarzyć, gdyby świat miał doświadczyć kolejnej raptownej zmiany *stopy wzrostu* porównywanej pod względem skali ze zmianami, jakie pociągnęły za sobą rewolucja neolityczna i rewolucja przemysłowa. Ekonomista Robin Hanson szacuje — opierając się na historycznych danych ekonomicznych i demograficznych — czas podwojenia rozmiarów światowej gospodarki dla plejstoceń-

skich społeczności myśliwsko-zbierackich na 224 tysiące lat, dla społeczności rolniczych na 909 lat i dla społeczeństw przemysłowych na 6,3 roku[3] (w modelu Hansona obecna epoka jest mieszaniną rolniczych i przemysłowych faz wzrostu — rozmiary gospodarki światowej jako całości nie podwajają się jeszcze co 6,3 roku). Gdyby miało dojść do kolejnego takiego przejścia do odmiennej fazy wzrostu i gdyby miało być to przejście porównywalne pod względem skali do dwóch poprzednich, doszłoby w efekcie do nowych warunków wzrostu, w których rozmiary gospodarki światowej ulegałyby podwojeniu mniej więcej co dwa tygodnie.

Taka stopa wzrostu ze współczesnej perspektywy może się wydawać iście fantastyczna. Obserwatorzy we wcześniejszych epokach musieli uważać za równie absurdalne podejrzenia, że pewnego dnia rozmiary światowej gospodarki będą się podwajać kilkakrotnie w czasie ludzkiego życia. A mimo to te wyjątkowe warunki są dziś naszym chlebem powszednim.

Koncepcja nadchodzącej technologicznej osobliwości została przez ostatnie lata szeroko spopularyzowana, poczynając od wpływowego eseju Vernora Vinge'a, którego następstwem były publikacje Raya Kurzweila i innych[4]. Jednakże termin „osobliwość" używany był myląco w wielu całkowicie odmiennych znaczeniach i dziś otacza go aura iście piekielnych (choć niemal tysiącletnich) konotacji techniczno-utopijnych[5]. Ponieważ większość tych znaczeń i konotacji nie ma dla naszego wywodu żadnego znaczenia, możemy osiągnąć większą klarowność, pozbywając się terminu „osobliwość" na rzecz terminologii precyzyjniejszej.

Związana z osobliwością koncepcja, która nas tutaj interesuje, to możliwość *eksplozji inteligencji*, a zwłaszcza perspektywa rozwoju superinteligentnych maszyn. Być może są tacy, których wykresy wzrostu przypominające te na rysunku 1. przekonują, że zanosi się na kolejną drastyczną zmianę tempa wzrostu, porównywalną z rewolucją neolityczną lub przemysłową. Ci ludzie mogą pomyśleć, że trudno jest opracować scenariusz, w którym czas podwajania się rozmiarów światowej gospodarki skraca się do zaledwie dwóch tygodni, jeśli nie uwzględni się w nim stworzenia umysłów znacznie szybszych i działających znacznie efektywniej niż znane nam mózgi biologiczne. Jednakże argumenty na rzecz poważnego rozważenia perspektyw rewolucji myślących maszyn nie muszą się opierać jedynie na ćwiczeniach z konstruowania odpowiedniego modelu regresji ani ekstrapolacji przeszłych trendów wzrostu gospodarczego. Jak się przekonamy, są poważniejsze przyczyny, by wziąć to pod uwagę.




**Rysunek 1.** Długookresowa historia wzrostu światowego PKB. Ukazana na skali liniowej historia światowej gospodarki wygląda jak płaska linia przytulona do osi x, która nagle szybuje niemal pionowo w górę. (a) Nawet kiedy skupimy się na ostatnich 10 tysiącach lat, wykres zasadniczo pozostaje ten sam: linia trendu skręca jednokrotnie pod kątem niemal 90 stopni ku górze. (b) Dopiero w ciągu ostatnich mniej więcej stu lat krzywa uniosła się dostrzegalnie ponad poziom zerowy (różne linie na wykresie odpowiadają różnym zbiorom danych, które pozwalają uzyskać lekko różniące się szacunki[6])

## Wielkie oczekiwania

Pojawienia się maszyn, które dorównają ludziom pod względem inteligencji ogólnej — a więc posiądą zdrowy rozsądek i faktyczną zdolność do uczenia się, wnioskowania i planowania, pozwalające rozwiązywać złożone problemy, które wymagają przetwarzania danych w szeregu obszarów zarówno abstrakcyjnych, jak i związanych z zagadnieniami życia codziennego — oczekiwano od chwili wynalezienia komputerów w latach 40. XX wieku. W tamtym czasie skonstruowania takich maszyn spodziewano

się w perspektywie kolejnych dwudziestu lat[7]. Od tamtego okresu oczekiwana data pojawienia się tego rodzaju maszyn była przesuwana co roku o rok. A zatem dziś futuryści rozważający możliwość pojawienia się sztucznej inteligencji nadal są często zdania, że od inteligentnych maszyn dzieli nas kilka dekad[8].

Dwie dekady to ulubiony przedział czasowy tych, którzy wieszczą radykalną zmianę — dość krótki, by zwrócić uwagę opinii publicznej i skłonić ją do uznania tematu za wart przedyskutowania, a mimo to wystarczająco długi, by umożliwić przypuszczenie, że w międzyczasie dojdzie do szeregu przełomów, które dziś trudno sobie wyobrazić, lecz które jednak mogą nastąpić. Porównajmy to z krótszymi przedziałami czasowymi: większość technologii, które będą wywierać znaczący wpływ na świat za pięć czy dziesięć lat, już w tej chwili jest w ograniczonym zakresie wykorzystywana, natomiast technologie, które przekształcą świat za mniej niż piętnaście lat, prawdopodobnie istnieją już w postaci prototypów rozwijanych w laboratoriach. Dwadzieścia lat to często również okres, w którym prawdopodobnie zakończy się kariera prognosty, co ogranicza potencjalnie negatywny wpływ śmiałych, lecz nietrafionych przewidywań na jego reputację.

Jednakże z faktu, że pewne osoby przewidywały w przeszłości zbyt rychłe pojawienie się sztucznej inteligencji, nie wynika jeszcze, że istnienie sztucznej inteligencji nie jest możliwe czy też nigdy nie zostanie ona opracowana[9]. Głównym powodem, dla którego dotychczasowy postęp był wolniejszy niż oczekiwano, jest to, że techniczne przeszkody stojące na drodze do skonstruowania inteligentnych maszyn okazały się większe, niż przewidywali to pionierzy. Otwarte pozostaje pytanie, jak duże są to przeszkody i ile nas jeszcze dzieli od ich przezwyciężenia. Czasem problem, który początkowo wydaje się beznadziejnie skomplikowany, okazuje się mieć zaskakująco proste rozwiązanie (choć częstsza jest prawdopodobnie sytuacja odwrotna).

W następnym rozdziale przyjrzymy się rozmaitym scenariuszom, które mogą doprowadzić do powstania maszyn dorównujących inteligencją człowiekowi. Zaznaczmy jednak już na wstępie, że niezależnie od tego, ile przystanków dzieli sytuację obecną od powstania sztucznej inteligencji dorównującej ludzkiej, jej powstanie nie jest punktem docelowym. Następnym przystankiem, leżącym w niedalekiej odległości od poprzedniego, jest sztuczna inteligencja przewyższająca ludzką. Pociąg niekoniecznie się zatrzyma czy choćby zwolni na stacji Ludzieszyn. Bardziej prawdopodobne jest to, że przemknie przez nią ze świstem.

Matematyk Irvin John Good, który pracował jako główny statystyk w zespole kryptologów Alana Turinga podczas II wojny światowej, prawdopodobnie jako pierwszy wyłożył kluczowe aspekty tego scenariusza. W często przywoływanym fragmencie z 1965 roku napisał:

> Zdefiniujmy maszynę ultrainteligentną jako maszynę, która dalece przewyższa na polu wszelkich aktywności intelektualnych dowolnego, choćby najinteligentniejszego człowieka. Ponieważ projektowanie maszyn jest jedną z tych aktywności intelektualnych, maszyna ultrainteligentna potrafi projektować jeszcze lepsze maszyny; dojdzie zatem bez wątpienia do „eksplozji inteligencji", w wyniku której inteligencja człowieka pozostanie daleko w tyle. A zatem pierwsza maszyna ultrainteligentna będzie ostatnim wynalazkiem, którego człowiek kiedykolwiek dokona — zakładając, że maszyna ta okaże się wystarczająco potulna, by powiedzieć nam, jak ją utrzymać pod kontrolą[10].

Może się teraz wydawać oczywiste, że z taką eksplozją inteligencji wiąże się poważne zagrożenie egzystencjalne i że z tego względu tę perspektywę należałoby rozważyć z największą powagą, nawet gdybyśmy mieli pewność (a jej nie mamy), że prawdopodobieństwo jej zaistnienia jest naprawdę niewysokie. Jednakże pionierzy sztucznej inteligencji — pomimo swojego przekonania o nieuchronności sztucznej inteligencji dorównującej ludzkiej — raczej nie rozważali prawdopodobieństwa zaistnienia sztucznej inteligencji przewyższającej ludzką. Tak jakby ich zdolność snucia domysłów całkowicie się wyczerpała po zrodzeniu radykalnej koncepcji maszyn dorównujących człowiekowi inteligencją, uniemożliwiając im dostrzeżenie jej naturalnej konsekwencji: koncepcji maszyn, które z czasem pod względem inteligencji przewyższą ludzi.

Większość pionierów sztucznej inteligencji (SI) nie przyjmuje do wiadomości, że podjęte przez nich przedsięwzięcie może się wiązać z ryzykiem[11]. Nie starają się nawet pozorować zainteresowania — nie wspominając już o poważnym namyśle — jakimikolwiek obawami o bezpieczeństwo czy wątpliwościami etycznymi związanymi ze stworzeniem sztucznych umysłów lub potencjalnych komputerowych władców; ta luka zdumiewa nawet na tle niezbyt imponujących standardów tej epoki dotyczących krytycznej oceny potencjału techniki[12]. Musimy mieć nadzieję, że do momentu, w którym to przedsięwzięcie rzeczywiście przyniesie owoce, posiądziemy nie tylko techniczną biegłość umożliwiającą nam zapoczątkowanie eksplozji inteligencji, ale i wyższy poziom zdolności jej kontrolowania, co może okazać się niezbędne, byśmy uszli z tej detonacji z życiem.

Zanim jednak zajmiemy się tym, co przed nami, przyda się szybki rzut oka na historię inteligentnych maszyn aż do czasów współczesnych.

## Okresy nadziei i rozczarowań

Latem 1956 roku w Dartmouth College dziesięciu naukowców zainteresowanych zagadnieniami sieci neuronowych, teorią automatów i badaniami nad inteligencją zjechało się na sześciotygodniowe warsztaty. Dartmouth Summer Research Project on Artificial Intelligence uznawany jest często za wydarzenie, które zapoczątkowało prace nad sztuczną inteligencją jako dziedzinę badań naukowych. Wielu uczestników wspomnianych warsztatów zostało później uznanych za ojców tej dziedziny. Optymistyczne poglądy dominujące wśród uczestników warsztatów odzwierciedla wniosek przedstawiony Fundacji Rockefellera, która sfinansowała to wydarzenie:

> Proponujemy przeprowadzenie dwumiesięcznego warsztatu badawczego dla dziesięciu badaczy zagadnień sztucznej inteligencji (...). Warsztat zostanie przeprowadzony przy założeniu, że każdy aspekt uczenia się lub dowolnej innej cechy inteligencji może zostać z zasady opisany tak precyzyjnie, że możliwe będzie zbudowanie maszyny do jego zasymulowania. Zostanie podjęta próba zrozumienia, jak zbudować maszyny posługujące się językiem naturalnym, formułujące idee i koncepcje abstrakcyjne, rozwiązujące problemy z gatunku tych zarezerwowanych dziś dla ludzi i doskonalące same siebie. Sądzimy, że można dokonać znaczących postępów w jednej lub kilku z tych dziedzin, jeśli starannie wybrana grupa naukowców podejmie nad nimi wspólną pracę podczas lata.

W ciągu sześciu dekad, które upłynęły od tego śmiałego początku, badania nad sztuczną inteligencją przechodziły naprzemiennie okresy wzmożonego zainteresowania i rozdmuchanych oczekiwań oraz niepowodzeń i rozczarowań.

Pierwszy okres ekscytacji, zapoczątkowany spotkaniem w Dartmouth, został później opisany przez Johna McCarthy'ego (głównego organizatora tego wydarzenia) jako era „Mamo, popatrz, wcale się nie trzymam!". W tym początkowym okresie badacze budowali systemy zaprojektowane z myślą o obaleniu twierdzeń w rodzaju „Żadna maszyna nigdy nie zdoła zrobić *X*!". W tamtych czasach tego rodzaju sceptyczne twierdzenia były na porządku dziennym. Aby się im przeciwstawić, badacze SI tworzyli małe systemy, którym udawało się zrobić *X* w „mikroświecie" (w dobrze zdefiniowanej, ograniczonej dziedzinie umożliwiającej zademonstrowanie zredukowanej wersji danego działania), dowodząc w ten sposób słuszności

danej koncepcji i pokazując, że zasadniczo działanie *X* może zostać wykonane przez maszynę. Jeden z tego rodzaju wczesnych systemów, program komputerowy Logic Theorist, był w stanie dowieść większości twierdzeń zamieszczonych w drugim rozdziale książki *Principia Mathematica* autorstwa Alfreda Northa Whiteheada i Bertranda Russella, a nawet zaproponować jeden dowód, który był znacznie bardziej elegancki niż oryginalny, i tym samym obalił przekonanie, że maszyny mogą „myśleć jedynie numerycznie", oraz zademonstrował, że maszyny są również zdolne przeprowadzić rozumowanie dedukcyjne i wymyślać dowody logiczne[13]. Program będący jego następcą, General Problem Solver, potrafił z zasady rozwiązywać szeroki zakres formalnie opisanych problemów[14]. Napisano również programy, które rozwikływały problemy rachunkowe na poziomie zajęć pierwszego roku college'u, dostrzegały wizualne analogie w zadaniach podobnych do tych, które zamieszcza się w niektórych testach na inteligencję, oraz znajdowały rozstrzygnięcia prostych problemów algebry opisowej[15]. Robot Shakey (nazwany tak ze względu na swoją skłonność do drżenia podczas wykonywania operacji) zaprezentował, w jaki sposób rozumowanie logiczne może zostać zintegrowane z postrzeganiem i wykorzystane do zaplanowania fizycznego działania i sterowania nim[16]. Program ELIZA pokazał, że komputer może stać się uosobieniem psychoterapeuty rogeriańskiego[17]. W połowie lat 70. program SHRDLU zademonstrował, w jaki sposób zasymulowane robotyczne ramię w zasymulowanym świecie geometrycznych klocków może wypełniać polecenia i odpowiadać na pytania zadane po angielsku przez użytkownika korzystającego z klawiatury[18]. W późniejszych dekadach miały powstać systemy, które dowiodły, że maszyny potrafią tworzyć muzykę w stylu rozmaitych kompozytorów klasycznych, przewyższyć mniej doświadczonych lekarzy w wykonywaniu pewnych zadań z zakresu diagnostyki klinicznej, autonomicznie prowadzić samochody i dokonywać wynalazków nadających się do objęcia ochroną patentową[19]. Stworzono nawet taki system SI, który umiał wymyślać oryginalne dowcipy[20]. (Nie to, żeby cechował się jakimś szczególnie wyrafinowanym poczuciem humoru, lecz ponoć dzieci zgodnie twierdziły, że jego żarty są śmieszne).

Trudne jednak okazało się rozciągnięcie metod, które pozwoliły osiągnąć sukcesy wczesnym wersjom demonstracyjnym, na szerszy zakres problemów lub na problemy trudniejsze. Jednym z powodów tej sytuacji jest „eksplozja kombinatoryczna" możliwości, które muszą zostać zbadane przez metody opierające się na mechanizmach wyszukiwania wyczerpującego. Takie metody dobrze się sprawdzają w odniesieniu do prostych problemów, lecz zawodzą, gdy sprawy zaczynają się komplikować. Dla

przykładu, aby dowieść twierdzenia, którego dowód zawiera się w pięciu liniach w systemie dedukcyjnym z jedną regułą wnioskowania i pięcioma aksjomatami, wystarczy po prostu wyliczyć 3125 możliwych kombinacji i sprawdzić każdą z nich, by stwierdzić, która daje oczekiwany wynik. Wyszukiwanie wyczerpujące sprawdza się również w przypadku dowodów zamykających się w sześciu czy siedmiu liniach. Kiedy jednak zadanie staje się bardziej skomplikowane, metody wyszukiwania wyczerpującego napotykają na problemy. Dowiedzenie twierdzenia, którego dowód mieści się w pięćdziesięciu liniach, nie trwa dziesięciokrotnie dłużej niż dowiedzenie twierdzenia, którego dowód mieści się w pięciu liniach. Jeśli korzysta się z wyszukiwania wyczerpującego, wymaga to przeanalizowania $5^{50} \approx 8{,}9 \times 10^{34}$ możliwych sekwencji — czego nie da się obliczyć nawet za pomocą najszybszych superkomputerów.

Do przezwyciężenia eksplozji kombinatorycznej potrzebne są algorytmy wykorzystujące struktury z dziedziny docelowej i zdobytą wcześniej wiedzę poprzez zastosowanie przeszukiwania heurystycznego, planowania i elastycznych reprezentacji abstrakcyjnych — tego rodzaju mechanizmy we wczesnych systemach SI były rozwinięte w stopniu znikomym. Potencjał tych wczesnych systemów podkopywały również marne metody radzenia sobie z niepewnością, poleganie na kruchych i nieugruntowanych reprezentacjach symbolicznych, niedostatek danych oraz poważne ograniczenia sprzętowe w zakresie pojemności pamięci i szybkości przetwarzania danych. W połowie lat 70. świadomość tych problemów stawała się coraz większa. Zrozumienie, że wielu projektom SI nigdy nie uda się spełnić pokładanych w nich oczekiwań, doprowadziło do nadejścia pierwszej „zimy SI": okresu redukcji wydatków, podczas którego ograniczeniu finansowania tego rodzaju projektów towarzyszył rosnący sceptycyzm. Sztuczna inteligencja wyszła z mody.

Nowa wiosna rozpoczęła się z początkiem lat 80., kiedy to Japończycy uruchomili swój projekt systemów komputerowych piątej generacji, hojnie sfinansowane partnerstwo prywatno-publiczne, którego celem było dokonanie ogromnego skoku technicznego poprzez rozwój potężnych systemów bazujących na architekturze przetwarzania równoległego, które miały się stać platformą systemów sztucznej inteligencji. Doszło do tego w szczycie fascynacji japońskim „powojennym cudem gospodarczym" — okresem, w którym to państwa zachodnie i liderzy biznesu starali się odgadnąć przepis Japończyków na sukces gospodarczy, mając nadzieję na powtórzenie tej magicznej sztuczki na własnym podwórku. Kiedy Japonia postanowiła zainwestować znaczne kwoty w rozwój sztucznej inteligencji, kilka innych krajów poszło w jej ślady.

W kolejnych latach doszło do upowszechnienia na szeroką skalę **systemów eksperckich**. Zaprojektowane jako narzędzia wspierające decydentów, systemy eksperckie były programami opartymi na regułach, które wyciągały proste wnioski z bazy wiedzy zawierającej fakty uzyskane od specjalistów w danej dziedzinie i z mozołem wklepane ręcznie do systemu w języku formalnym. Zbudowano setki tego rodzaju systemów eksperckich, ale mniejsze systemy okazały się mało przydatne, a większe bardzo kosztowne, gdy szło o ich rozwój, weryfikację i aktualizację, przy czym z zasady były bardzo nieporęczne i kłopotliwe w użyciu. Kupowanie samodzielnego komputera tylko po to, by uruchomić na nim jeden program, było niepraktyczne. Pod koniec lat 80. również ten okres wzrostu dobiegł końca.

Projektowi piątej generacji nie udało się osiągnąć zakładanych celów, podobnie jak analogicznym do niego projektom uruchomionym w Stanach Zjednoczonych i Europie. Nadeszła druga zima SI. W tamtym momencie krytycy mogli słusznie ubolewać nad „historią dotychczasowych badań nad sztuczną inteligencją, które nieodmiennie mogły się pochwalić jedynie bardzo ograniczonymi sukcesami w ściśle określonych dziedzinach, a po tych sukcesach następowało natychmiastowe fiasko projektów zmierzających do osiągnięcia szerzej zakreślonych celów, na których realizację te początkowe sukcesy wydawały się dawać nadzieję”[21]. Prywatni inwestorzy zaczęli unikać jakichkolwiek przedsięwzięć realizowanych pod hasłem „sztucznej inteligencji”. Nawet wśród naukowców i ich fundatorów „SI” stało się niepożądanym epitetem[22].

Prace techniczne posuwały się jednak w szybkim tempie naprzód i w latach 90. znów rozpoczęła się odwilż. Ziarno optymizmu zostało zasiane dzięki wprowadzeniu nowych technik, które wydawały się alternatywą do tradycyjnego paradygmatu logicznego (określanego czasem mianem „starej dobrej sztucznej inteligencji”, Good Old Fashioned Artificial Intelligence, w skrócie GOFAI) — ten koncentrował się na wysokopoziomowej manipulacji symbolicznej, a jego kulminacją były systemy eksperckie lat 80. Cieszące się świeżo zdobytą popularnością techniki sieci neuronowych i algorytmów genetycznych obiecywały przezwyciężenie niektórych problemów podejścia GOFAI, zwłaszcza „kruchości” cechującej tradycyjne programy SI (wyrzucające zazwyczaj kompletne bzdury, gdy programiści poczynili choćby jedno nie do końca prawidłowe założenie). Nowe techniki chlubiły się działaniem bardziej organicznym. Dla przykładu: sieci neuronowe wykazywały własność „płynnego obniżania efektywności”: niewielkie uszkodzenie sieci neuronowej owocowało zazwyczaj niewielkim spadkiem jej wydajności, a nie całkowitym krachem systemu. Co jeszcze

istotniejsze, sieci neuronowe potrafiły się uczyć na podstawie własnych doświadczeń, znajdując naturalne sposoby generalizowania na bazie przykładów i wyszukiwania ukrytych wzorców statystycznych w danych wejściowych[23]. Dzięki tej cesze sieci trafnie rozpoznawały trendy i sprawnie klasyfikowały problemy. Dla przykładu: szkoląc sieć neuronową na podstawie zestawów danych będących pomiarami sonarów, można było nauczyć ją rozpoznawania profili akustycznych łodzi podwodnych, min i fauny morskiej z trafnością nieosiągalną dla ludzkich specjalistów — i można było tego dokonać bez konieczności określania przez kogokolwiek z góry dokładnych definicji kategorii czy też konkretnych wag przypisywanych rozmaitym cechom.

O ile proste modele sieci neuronowych były znane już w późnych latach 50., o tyle dziedzina ta przeżyła renesans po wprowadzeniu algorytmu wstecznej propagacji błędu, który umożliwił szkolenie wielowarstwowych sieci neuronowych[24]. Takie wielowarstwowe sieci, które mają jedną lub więcej pośrednich („ukrytych") warstw neuronów pomiędzy warstwami wejściową i wyjściową, potrafią opanować działania ze znacznie szerszego zakresu niż ich prostsze poprzedniczki[25]. W połączeniu z coraz wydajniejszymi komputerami, które stawały się coraz łatwiej dostępne, te udoskonalenia algorytmiczne umożliwiły inżynierom zbudowanie sieci neuronowych wystarczająco dobrych, by mogły znaleźć wiele praktycznych zastosowań.

Przypominające pod względem własności ludzki mózg sieci neuronowe wypadały korzystnie w zestawieniu ze sztywnymi, bezwzględnie logicznymi, lecz przy tym kruchymi tradycyjnymi systemami GOFAI opartymi na regułach — dość korzystnie, by przyczyniły się do powstania nowego -izmu: **koneksjonizmu**, który podkreślał istotność przetwarzania masowo równoległego. Od tego czasu opublikowano ponad 150 tysięcy artykułów naukowych na temat sztucznych sieci neuronowych będących nadal istotnym podejściem do kwestii uczenia się maszyn.

Metody oparte na ewolucji, takie jak algorytmy genetyczne i programowanie genetyczne, stanowią kolejne podejście, którego pojawienie się przyczyniło się do zakończenia drugiej zimy AI. Wywarło być może mniejszy wpływ na badania naukowe niż sieci neuronowe, lecz zostało szeroko spopularyzowane. W modelach ewolucyjnych utrzymywana jest populacja rozwiązań „kandydujących" (którymi mogą być struktury danych lub programy), a nowe rozwiązania tworzone są losowo poprzez mutacje lub rekombinację wariantów w istniejącej populacji. Od czasu do czasu populacja jest redukowana poprzez zastosowanie kryterium selekcji (funkcja dostosowania darwinowskiego), które pozwala tylko lepszym kandydatom

przetrwać do następnego pokolenia. Po przeprowadzeniu tysięcy iteracji w kolejnych pokoleniach średnia jakość rozwiązań w puli kandydatów stopniowo wzrasta. Tego rodzaju algorytm, jeśli zadziała, pozwala na uzyskanie wydajnych rozwiązań bardzo szerokiego zakresu problemów — rozwiązań, które mogą być uderzająco nowatorskie i sprzeczne z intuicją, a które często bardziej przypominają naturalne struktury niż cokolwiek, co mogłoby zostać zaprojektowane przez inżyniera. Z zasady można to osiągnąć bez znaczącej interwencji po stronie człowieka, jeśli pominąć początkowe określenie funkcji dopasowania, która jest często bardzo prosta. W praktyce jednak doprowadzenie do tego, by metody ewolucyjne przynosiły oczekiwane rezultaty, wymaga umiejętności i pomysłowości, zwłaszcza gdy chodzi o opracowanie właściwego formatu reprezentacji danych. Bez skutecznej metody zakodowania rozwiązań kandydujących (języka genetycznego, który odpowiada ukrytej strukturze dziedziny docelowej), wyszukiwanie ewolucyjne ma skłonność do meandrowania w nieskończoność w rozległej przestrzeni poszukiwań lub utykania w lokalnym optimum. Nawet gdy zostanie znaleziony dobry format reprezentacji, ewolucja wymaga dużej mocy obliczeniowej i często pada ofiarą eksplozji kombinatorycznej.

Sieci neuronowe i algorytmy genetyczne są przykładami metod, które wywoływały podniecenie w latach 90., gdyż wydawało się, że stanowią alternatywę do trwającego w zastoju paradygmatu GOFAI. Moją intencją w tym miejscu nie jest jednak wyśpiewywanie peanów na cześć tych dwóch metod ani wynoszenie ich ponad wiele innych technik uczenia się maszyn. Tak naprawdę jednym z głównych teoretycznych osiągnięć ostatnich dwudziestu lat było dobitniejsze uzmysłowienie nam, jak pozornie odmienne techniki mogą zostać ujęte jako odmienne przypadki w ramach wspólnego modelu matematycznego. Dla przykładu: wiele typów sztucznych sieci neuronowych można uznać za klasyfikatory, które wykonują pewien szczególny rodzaj obliczeń statystycznych (estymacja metodą największej wiarygodności)[26]. Ta perspektywa pozwala porównać sieci neuronowe z szerszą klasą algorytmów klasyfikatorów uczenia się na przykładach, między innymi z: drzewami decyzyjnymi, modelami regresji logistycznej, metodą wektorów nośnych, naiwnym klasyfikatorem bayesowskim i metodą *k*-najbliższych sąsiadów[27]. W analogiczny sposób algorytmy genetyczne można uznać za postać sterowania stochastycznego, co znów jest podzbiorem szerszej klasy algorytmów optymalizacyjnych. Każdy z tych algorytmów pozwalających budować klasyfikatory lub przeszukiwać przestrzeń rozwiązań ma własne charakterystyczne słabe i mocne strony, które można badać matematycznie. Algorytmy różnią się, gdy idzie o wymaganą

pamięć i moc obliczeniową, przyjmowane z góry ukierunkowanie indukcyjne, łatwość włączenia treści z zewnątrz oraz przejrzystości mechanizmów ich działania dla analizującego je człowieka.

Mętna otoczka uczenia maszynowego i twórczego rozwiązywania problemów kryje zatem zestaw dobrze określonych matematycznie kompromisów. Ideałem jest perfekcyjny sprawca bayesowski — taki, który czyni probabilitycznie optymalny użytek z dostępnych informacji. Ten ideał jest nieosiągalny, ponieważ wymaga zbyt dużej mocy obliczeniowej, by mógł zostać wdrożony na jakimkolwiek fizycznym komputerze (patrz: ramka 1.). W związku z tym sztuczną inteligencję można uznać za pościg za ułatwieniami — metodami sukcesywnego przybliżania się do bayesowskiego ideału poprzez poświęcanie w jakiejś mierze optymalności lub ogólności przy jednoczesnym zachowaniu ich w takim stopniu, by możliwe było uzyskanie wysokiej wydajności w faktycznej dziedzinie zainteresowania.

Odzwierciedlenie tego obrazu można dostrzec w przeprowadzonych przez kilka ostatnich dekad pracach dotyczących modeli prawdopodobieństwa wykorzystujących grafy, takich jak sieci bayesowskie. Sieci bayesowskie stanowią zwięzłą metodę reprezentacji probabilistycznych i warunkowych relacji niezależności obowiązujących w pewnej konkretnej domenie. (Wykorzystanie takich relacji niezależności jest kluczowe dla przezwyciężenia eksplozji kombinatorycznej, która stwarza takie same problemy przy wnioskowaniu probabilistycznym, jak przy dedukcji logicznej). Pozwalają również uzyskać istotny wgląd w koncepcję przyczynowości[28].

Jedną z zalet odnoszenia problemów uczenia się z konkretnej dziedziny do generalnych problemów wnioskowania bayesowskiego jest to, że nowe algorytmy usprawniające wnioskowanie bayesowskie przynoszą również natychmiastową poprawę w wielu innych obszarach. Dla przykładu: postępy w rozwoju metody Monte Carlo znajdują bezpośrednie zastosowanie w obrazowaniu komputerowym, robotyce i genetyce obliczeniowej. Kolejnym plusem jest to, że badacze zajmujący się wieloma różnymi dyscyplinami mogą z większą łatwością dzielić się swoimi odkryciami. Badania nad statystykami bayesowskimi i modelami opartymi na grafach prowadzone są jednocześnie na wielu polach, włączając w to uczenie maszynowe, fizykę statystyczną, bioinformatykę, optymalizację kombinatoryczną i teorię komunikacji[29]. Znaczna część ostatnich postępów w dziedzinie uczenia maszynowego jest wynikiem włączenia formalnych wyników uzyskanych w innych dziedzinach badań naukowych. (Programy uczenia maszynowego skorzystały również ogromnie na pojawieniu się szybszych komputerów i większej dostępności dużych zestawów danych).

## Ramka 1. Optymalny sprawca bayesowski

Idealny sprawca bayesowski wychodzi od „rozkładu prawdopodobieństwa a priori" — funkcji, która przypisuje prawdopodobieństwo każdemu z „możliwych światów" (na przykład każdej maksymalnie skonkretyzowanej postaci, którą ów świat może przybrać)[30]. Przyjęte a priori założenie to ukierunkowanie indukcyjne, że na przykład prostszym światom przypisywane jest wyższe prawdopodobieństwo (jedną z możliwych metod formalnego zdefiniowania prostoty możliwego świata jest ujęcie jej w kategoriach „złożoności Kołmogorowa" — miary bazującej na długości najkrótszego programu komputerowego, który pozwala uzyskać pełny opis świata)[31]. W tym uprzednim założeniu mieści się również całość wiedzy ogólnej, którą programiści chcieliby obdarzyć sprawcę.

Kiedy sprawca otrzymuje nowe informacje ze swoich czujników, aktualizuje rozkład prawdopodobieństwa, uwarunkowując ten rozkład nowymi informacjami zgodnie z twierdzeniem Bayesa[32]. Uwarunkowywanie funkcji prawdopodobieństwa jest operacją matematyczną, która przypisuje nowe, zerowe prawdopodobieństwo tym światom, które są sprzeczne z otrzymaną informacją, i renormalizuje rozkład prawdopodobieństwa na pozostałe możliwe światy. Wynikiem jest „rozkład prawdopodobieństwa *a posteriori*" (który sprawca może wykorzystać w następnym kroku jako nowy rozkład *a priori*). W miarę jak sprawca dokonuje obserwacji, coraz bardziej prawdopodobne staje się zaistnienie coraz mniejszej liczby możliwych światów nadal zgodnych z zebranymi danymi z obserwacji — a wśród tych możliwych światów światy prostsze są zawsze bardziej prawdopodobne.

Metaforycznie możemy myśleć o prawdopodobieństwie jako o piasku rozsypanym na dużej kartce papieru. Kartka dzielona jest na obszary rozmaitych rozmiarów, przy czym każdy obszar odpowiada jednemu możliwemu światu, a większe obszary to prostsze możliwe światy. Wyobraźmy sobie warstwę piasku równej grubości pokrywającą równo całą kartkę — to jest nasz rozkład prawdopodobieństwa *a priori*. Za każdym razem, gdy dokonywana jest obserwacja wykluczająca jeden z możliwych światów, usuwamy piasek z odpowiedniego obszaru na kartce i rozdzielamy go równo na obszary, które są nadal brane pod uwagę. W ten sposób całkowita ilość piasku na kartce nigdy nie ulega zmianie, po prostu koncentruje się on na coraz mniejszej liczbie obszarów w miarę gromadzenia kolejnych danych z obserwacji. Oto koncepcja uczenia się w jej najczystszej formie. (Aby obliczyć prawdopodobieństwo *hipotezy*, mierzymy po prostu ilość piasku we wszystkich obszarach odpowiadających możliwym światom, dla których hipoteza jest prawdziwa).

Dotychczas zdefiniowaliśmy regułę uczenia się. Aby uzyskać sprawcę, potrzebujemy jeszcze reguły decyzyjnej. W tym celu wyposażamy sprawcę w funkcję użyteczności, która przypisuje każdemu z możliwych światów pewną liczbę. Ta liczba reprezentuje pożądalność tego świata zgodnie z zasadniczymi preferencjami sprawcy. Teraz przy każdym kroku sprawca wybiera działanie o najwyższej oczekiwanej użyteczności[33] (aby znaleźć działanie o najwyższej oczekiwanej użyteczności, sprawca może wyliczyć wszystkie możliwe działania, a następnie rozkład prawdopodobieństwa warunkowego, biorąc pod uwagę podjęcie tego działania, a więc rozkład prawdopodobieństwa, który będzie wynikiem uwarunkowania obecnego rozkładu prawdopodobieństwa od obserwacji, że działanie właśnie zostało podjęte; wreszcie może obliczyć oczekiwaną wartość działania jako sumę wartości każdego możliwego świata przemnożoną przez prawdopodobieństwo warunkowe tego świata przy założeniu podjęcia tego działania[34]).

Reguła uczenia się i reguła decyzyjna razem wzięte określają dla sprawcy „pojęcie optymalności" (zasadniczo to samo „pojęcie optymalności" jest szeroko wykorzystywane w dziedzinie sztucznej inteligencji, teorii poznania, filozofii nauki, ekonomii i statystyce[35]). W rzeczywistości niemożliwe jest zbudowanie takiego sprawcy, ponieważ ze względu na wymaganą moc obliczeniową niemożliwe jest dokonanie wymaganych obliczeń. Każda próba skończy się fiaskiem z powodu eksplozji kombinatorycznej takiej samej jak ta, którą opisaliśmy przy okazji naszej dyskusji na temat GOFAI. Aby zrozumieć, dlaczego tak się dzieje, rozważmy jeden niewielki podzbiór wszystkich możliwych światów — te światy, które składają się z pojedynczego monitora komputerowego unoszącego się w bezkresnej próżni. Monitor ma tysiąc na tysiąc pikseli i każdy z nich nieprzerwanie albo się świeci, albo nie. Nawet ten podzbiór możliwych światów jest ogromnie wielki: $2^{(1000\times1000)}$ możliwych stanów monitora przewyższa pod względem liczby wszystkie obliczenia, które kiedykolwiek nastąpią w dającym się zaobserwować świecie. A zatem nie potrafimy nawet wymienić wszystkich możliwych światów tego małego podzbioru wszystkich możliwych światów, nie mówiąc już o wykonaniu bardziej skomplikowanych obliczeń dotyczących każdego z nich oddzielnie.

Pojęcia optymalności mogą stanowić przedmiot rozważań teoretycznych — nawet jeśli nie da się ich zrealizować fizycznie. Oferują nam punkt odniesienia pozwalający ocenić przybliżenia heurystyczne i czasem pomagają wywnioskować, co w niektórych konkretnych przypadkach zrobiłby optymalny sprawca. W rozdziale 12. zetkniemy się z pewnymi alternatywnymi pojęciami optymalności odnoszącymi się do sztucznych sprawców.

# Aktualny stan badań

Już teraz sztuczna inteligencja w wielu dziedzinach osiąga wyniki lepsze niż ludzie. W tabeli 1. analizujemy osiągnięcia komputerów grających w gry, dowodząc, że systemy SI pokonują w tej chwili mistrzów bardzo wielu gier[36].

**Tabela 1.** Systemy sztucznej inteligencji grające w gry

| | | |
|---|---|---|
| Warcaby | Przewyższa człowieka | Program szachowy Arthura Samuela napisany w 1952 roku, a następnie udoskonalony (w wersji z 1955 roku uwzględniono mechanizmy uczenia maszynowego) stał się pierwszym programem, który nauczył się grać w tę grę lepiej od swojego twórcy[37]. W 1994 roku program CHINOOK pokonał aktualnego mistrza szachowego; wówczas to po raz pierwszy program komputerowy wygrał oficjalne światowe mistrzostwa w grze opartej na umiejętnościach. W 2002 roku Jonathan Schaeffer i jego zespół „rozwiązują" szachy, czyli tworzą program, który zawsze wykonuje najlepszy możliwy ruch (łącząc przeszukiwanie algorytmem alfa-beta z bazą danych 39 bilionów pozycji w końcówce rozgrywki). Perfekcyjna gra po obu stronach doprowadziła do remisu[38]. |
| Backgammon | Przewyższa człowieka | 1979 rok: program do gry w backgammona o nazwie BGK autorstwa Hansa Berlinera pokonał mistrza świata; był to pierwszy program komputerowy, który pokonał (w pokazowym meczu) mistrza świata dowolnej gry, chociaż Berliner przypisał później to zwycięstwo szczęściu przy rzucie kośćmi[39].<br><br>1992 rok: program do gry w backgammona o nazwie TD-Gammon autorstwa Gerry'ego Tesaura osiągnął umiejętności mistrzowskie, wykorzystując algorytm uczenia się ze wzmocnieniem TD-Learing i rozgrywając szereg partii przeciwko samemu sobie w celu samodoskonalenia się[40].<br><br>Od tamtej pory programy grające w backgammona dalece prześcignęły najlepszych ludzkich graczy[41]. |
| Traveller TCS | Przewyższa człowieka we współpracy z człowiekiem[42] | Zarówno w 1981, jak i w 1982 roku program o nazwie Eurisko autorstwa Douglasa Lenata wygrał mistrzostwa Stanów Zjednoczonych w Traveller TCS (futurystycznej morskiej grze wojennej), przyczyniając się do zmiany zasad, której celem było zablokowanie jego niekonwencjonalnych strategii[43]. Eurisko ma wbudowane heurystyki umożliwiające projektowanie jego floty, a także heurystyki umożliwiające modyfikacje jego heurystyk. |
| Othello | Przewyższa człowieka | 1997 rok: program Logistello wygrał wszystkie rozgrywki w składającym się z sześciu partii meczu przeciwko mistrzowi świata Takeshiemu Murakamiemu[44]. |

**Tabela 1.** Systemy sztucznej inteligencji grające w gry — *ciąg dalszy*

| | | |
|---|---|---|
| Szachy | Przewyższa człowieka | 1997 rok: komputer Deep Blue pokonał mistrza świata, Garriego Kasparowa. Kasparow utrzymywał, że w niektórych ruchach komputera dostrzegł przebłyski prawdziwej inteligencji i kreatywności[45]. Od tamtej pory silniki szachowe stają się coraz doskonalsze[46]. |
| Krzyżówki | Poziom ekspercki | 1999 rok: program rozwiązujący krzyżówki o nazwie Proverb osiągnął lepsze wyniki niż przeciętny miłośnik krzyżówek[47]. |
| | | 2012 rok: program Dr. Fill stworzony przez Matta Ginsberga osiągnął wyniki w górnym kwartylu, konkurując z uczestnikami turnieju krzyżówkowego American Crossword Puzzle Tournament. (Wyniki Dr. Filla były nierówne. Rozwiązywał perfekcyjnie krzyżówki uznane przez ludzi za najtrudniejsze, lecz nie potrafił wykonać kilku niestandardowych łamigłówek obejmujących między innymi literowanie wstecz lub wpisywanie odpowiedzi po skosie[48]). |
| Scrabble | Przewyższa człowieka | Już w 2002 roku programy grające w Scrabble osiągały lepsze wyniki niż najlepsi gracze spośród ludzi[49]. |
| Brydż | Równy najlepszym | W 2005 roku programy grające w brydża kontraktowego zaczęły osiągać równie dobre wyniki jak najlepsi gracze w brydża[50]. |
| Jeopardy! | Przewyższa człowieka | 2010 rok: system komputerowy Watson wyprodukowany przez IBM pokonał dwóch mistrzów wszech czasów *Jeopardy!*: Kenna Jenningsa i Brada Ruttera[51]. *Jeopardy!* (w Polsce emitowany jako *Va banque*) to teleturniej wiedzy z pytaniami z rozmaitych dziedzin, między innymi z: historii, literatury, sportu, geografii, kultury popularnej i nauk ścisłych. Pytania prezentowane są w formie zagadek i często zawierają gry słów. |
| Poker | Różnie | Programy komputerowe grające w pokera nadal nieco ustępują najlepszym graczom w wariancie full-ring Texas Hold'em, ale w niektórych innych wariantach pokera osiągają wyniki lepsze niż ludzie[52]. |
| FreeCell | Przewyższa człowieka | Heurystyki powstałe w wyniku ewolucji z wykorzystaniem algorytmów genetycznych pozwoliły stworzyć program układający pasjansa FreeCell (który w swojej uogólnionej postaci jest NP-zupełny) umiejący pokonać plasujących się wysoko w rankingach graczy[53]. |
| Go | Bardzo wysoki poziom amatorski | W 2012 roku programy z serii Zen grające w go osiągnęły poziom 6 dana w szybkich grach (poziom bardzo silnego gracza grającego amatorsko), wykorzystując heurystykę MCTS i techniki uczenia maszynowego[54]. W ostatnich latach programy grające w go są udoskonalane w tempie mniej więcej 1 dana rocznie. Jeśli dalej w tym tempie będą doskonalone, być może w ciągu dekady uda im się pokonać mistrza świata. |

Te osiągnięcia dziś mogą nie wydawać się imponujące. Wynika to jednak wyłącznie z tego, że nasze wyobrażenie o tym, co jest imponujące, a co nie, zmienia się wraz z dokonującym się postępem. Dla przykładu:

biegłość w grze w szachy uznawano niegdyś za papierek lakmusowy ludzkiej inteligencji. W opinii kilku specjalistów z końca lat 50. „człowiek, który potrafiłby opracować skuteczną maszynę grającą w szachy, uznany zostałby za kogoś, komu udało się dotrzeć do samego rdzenia ludzkiego potencjału intelektualnego[55]". W tej chwili już się tak nie wydaje. Można się solidaryzować z Johnem McCarthym, który ubolewa: „Kiedy tylko coś zaczyna działać, przestaje się to określać mianem sztucznej inteligencji[56]".

W pewnym jednak istotnym sensie systemy SI dysponujące umiejętnością gry w szachy okazały się zwycięstwem mniej spektakularnym, niż wielu to sobie wyobrażało. Kiedyś zakładano, być może całkiem rozsądnie, że po to, by komputer mógł grać w szachy na poziomie mistrzowskim, należy wyposażyć go w wysoki poziom inteligencji *ogólnej*[57]. Można by na przykład pomyśleć, że mistrzowska gra w szachy wymaga umiejętności uczenia się pojęć abstrakcyjnych, inteligentnego planowania strategicznego, opracowywania elastycznych planów, nieustannego wysuwania oryginalnych wniosków logicznych, a może nawet modelowania sposobu myślenia przeciwnika. Tak nie jest. Okazało się, że możliwe jest zbudowanie świetnie sobie radzącego silnika szachowego na bazie specjalizowanego algorytmu[58]. Zaimplementowany na szybkich procesorach, które stały się dostępne pod koniec XX wieku, daje w efekcie grę na bardzo wysokim poziomie. Jednak SI zbudowane w ten sposób są wąskie. Grają w szachy, lecz nie potrafią robić nic innego[59].

W innych dziedzinach rozwiązania okazały się *bardziej* skomplikowane, niż początkowo zakładano, i postęp dokonywał się wolniej. Zajmującego się naukowo informatyką Donalda Knutha uderzyło to, że „SI udało się dotychczas zrobić zasadniczo wszystko, co wymaga »myślenia«, lecz nie udało im się zrobić większości tego, co ludzie i zwierzęta robią »bezmyślnie« — a co jakimś sposobem okazuje się znacznie trudniejsze!"[60]. Analiza obrazów, rozpoznawanie przedmiotów lub sterowanie zachowaniem robota wchodzącego w interakcje ze środowiskiem naturalnym stwarza znacznie więcej problemów. Mimo wszystko i w tej dziedzinie nastąpił spory postęp i dokonuje się on nadal, czemu sprzyja systematyczne ulepszanie sprzętu komputerowego.

Trudne okazało się również wyposażenie systemów SI w zdrowy rozsądek i zdolność rozumienia języka naturalnego. Sądzi się dzisiaj często, że osiągnięcie ludzkiego poziomu wykonywania tych zadań przez maszyny będzie tożsame z uzyskaniem sztucznej inteligencji, co oznacza, że trudności z rozwiązaniem tych problemów są zasadniczo równe trudnościom ze zbudowaniem maszyn dysponujących ludzką inteligencją ogólną[61]. Innymi słowy, gdyby komuś *udało się* zbudować system SI zdolny rozu-

mieć język naturalny równie dobrze jak dorosły człowiek, oznaczałoby to najprawdopodobniej, że albo już stworzył sztuczną inteligencję dorównującą ludzkiej, albo od jej stworzenia dzieli go zaledwie niewielki krok[62].

Umiejętność gry w szachy okazała się osiągalna przy wykorzystaniu bardzo prostego algorytmu. Kusi, by nabrać podejrzeń, że inne zdolności — takie jak ogólna zdolność rozumowania albo pewne kluczowe umiejętności, których wymaga programowanie — mogą być również osiągalne poprzez wykorzystanie jakiegoś zaskakująco prostego algorytmu. Fakt, że do najlepszych w danym momencie wyników dochodzi się poprzez wykorzystanie skomplikowanych mechanizmów, nie oznacza, że nie istnieje żaden prosty mechanizm pozwalający wykonać tę pracę równie dobrze lub lepiej. Może być po prostu tak, że nikt jeszcze nie znalazł prostszej alternatywy. Model Ptolemejski (z Ziemią w centrum okrążaną przez Słońce, Księżyc, planety i gwiazdy) odzwierciedlał stan wiedzy astronomicznej przez ponad tysiąc lat, a jego dokładność, gdy szło o przepowiadanie przyszłych zdarzeń, była systematycznie powiększana na przestrzeni wieków dzięki jego systematycznemu komplikowaniu polegającemu na dodawaniu kolejnych epicykli zakładanych ruchów ciał niebieskich. Później cały system został obalony przez teorię heliocentryczną Kopernika, która była prostsza i — choć dopiero po dopracowaniu jej przez Keplera — pozwalała na dokładniejsze przewidywania[63].

Metody sztucznej inteligencji są obecnie wykorzystywane w tak wielu obszarach, że nie miałoby sensu wyliczanie ich tutaj, ale wspomnienie choćby o wybranych da pewne wyobrażenie o mnogości zastosowań. Oprócz systemów SI potrafiących grać w gry, a wymienionych w tabeli 1., istnieją aparaty słuchowe wyposażone w algorytmy, które odsiewają szumy z otoczenia, systemy nawigacji wyświetlające mapy i prowadzące kierowców do wybranego punktu, systemy rekomendacji, które proponują książki i albumy muzyczne na podstawie wcześniejszych zakupów i ocen użytkownika, oraz systemy wspierające podejmowanie decyzji medycznych, które pomagają lekarzom diagnozować raka piersi, proponują plan leczenia i wspierają specjalistów w interpretacji elektrokardiogramów. Istnieją mechaniczne zwierzątka domowe i roboty sprzątające, roboty koszące trawniki, roboty ratownicze, chirurgiczne oraz ponad milion robotów przemysłowych[64]. Liczba pracujących na świecie robotów przekracza 10 milionów[65].

Współczesne systemy rozpoznawania mowy oparte na technikach statystycznych, takich jak ukryte modele Markowa, stały się wystarczająco dokładne dla zastosowań praktycznych (niektóre fragmenty tej książki zostały wstępnie napisane z pomocą programu rozpoznawania mowy).

Osobiści asystenci cyfrowi, tacy jak Siri autorstwa firmy Apple, reagują na polecenia głosowe, potrafią odpowiadać na proste pytania i wykonywać polecenia. Systemy rozpoznawania znaków (Optical Character Recognition, OCR) zarówno zapisanych ręcznie, jak i maszynowo, są wykorzystywane rutynowo w zastosowaniach takich jak sortowanie przesyłek pocztowych czy digitalizacji starych dokumentów[66].

Tłumaczenie maszynowe pozostaje niedoskonałe, lecz w wielu zastosowaniach jest wystarczająco dobre. Wcześniejsze systemy wykorzystywały podejście GOFAI opierające się na zaprogramowanych ręcznie zasadach gramatyki, które musiały zostać opracowane od zera dla każdego języka przez dysponujących odpowiednimi umiejętnościami lingwistów. Nowsze systemy wykorzystują techniki statystycznego uczenia maszynowego, które automatycznie budują modele statystyczne na podstawie zaobserwowanych wzorców użycia. System ustala parametry tych modeli, analizując dwujęzyczne korpusy tekstów. To podejście pozwala obyć się bez językoznawców — programiści tworzący te systemy nie muszą nawet mówić w języku, z którym pracują[67].

Systemy rozpoznawania twarzy zostały w ostatnich latach tak udoskonalone, by mogły zacząć być wykorzystywane na zautomatyzowanych przejściach granicznych w Europie i Australii. Amerykański Departament Stanu wykorzystuje w procedurach rozpatrywania wniosków wizowych system rozpoznawania twarzy dysponujący ponad 75 milionami fotografii. Systemy inwigilacji korzystają z coraz większej liczby wyrafinowanych technologii SI i eksploracji danych w analizie tekstów, głosu i nagrań wideo, które w ogromnych ilościach pobierane są ze światowych systemów komunikacji elektronicznej i przechowywane w gigantycznych centrach danych.

Programy umożliwiające dowodzenie twierdzeń i rozwiązywanie równań zostały dopracowane do tego stopnia, że dziś już właściwie nie uznaje się ich za systemy sztucznej inteligencji. Programy do rozwiązywania równań zostały włączone do programów wspierających obliczenia naukowe, takich jak Mathematica. Formalne metody weryfikacji, w tym zautomatyzowane programy dowodzenia twierdzeń, są rutynowo wykorzystywane przez producentów procesorów, by zweryfikować zachowanie projektowanych układów przed rozpoczęciem ich produkcji.

Amerykańskie służby wojskowe i wywiadowcze były prekursorami wykorzystania na szeroką skalę robotów zrzucających bomby, bezzałogowych samolotów szpiegowskich i bojowych oraz innych pojazdów bezzałogowych. W dalszym ciągu te maszyny zależą w dużej mierze od steru-

jących nimi zdalnie operatorów będących ludźmi, lecz prowadzone są prace zmierzające do powiększenia ich autonomii.

Obszarem, w którym odnotowano znaczące sukcesy, jest inteligentne harmonogramowanie. Narzędzie DART do zautomatyzowanego planowania logistyki i harmonogramowania zostało wykorzystane podczas operacji Pustynna Burza w 1991 roku z takim powodzeniem, że DARPA (amerykańska Defense Advanced Research Projects Agency — Agencja Zawansowanych Projektów Badawczych w Obszarze Obronności) utrzymuje, że ta jedna aplikacja z naddatkiem zwróciła trzydziestoletnie nakłady na rozwój systemów SI[68]. Systemy rezerwacji lotniczej wykorzystują zaawansowane algorytmy harmonogramowania i wyceny. Przedsiębiorstwa na szeroką skalę korzystają z technologii SI w systemach zarządzania stanami magazynowymi; wykorzystują również zautomatyzowane systemy rezerwacji telefonicznej oraz infolinie powiązane z programami rozpoznawania mowy, aby pokierować swoich nieszczęsnych klientów poprzez labirynt zawikłanych opcji menu.

Technologie SI leżą u podstaw wielu usług internetowych. Oprogramowanie kieruje światowym ruchem poczty elektronicznej i mimo tego, że spamerzy nieustannie modyfikują swoje strategie, by obejść stawiane przed nimi przeszkody, bayesowskie filtry antyspamowe w dużej mierze zdołały powstrzymać lawinę spamu. Oprogramowanie wykorzystujące komponenty sztucznej inteligencji odpowiada za automatyczne zatwierdzanie lub odrzucanie transakcji wykonywanych kartami kredytowymi i stale monitoruje operacje wykonywane na koncie, wypatrując oznak oszustwa. Systemy wyszukiwania informacji również w szerokim zakresie korzystają z uczenia maszynowego. Wyszukiwarka Google jest prawdopodobnie największym dotychczas zbudowanym systemem SI.

Należy jednak podkreślić, że linia demarkacyjna dzieląca systemy sztucznej inteligencji i oprogramowanie jako takie nie jest wyraźna. Niektóre z wymienionych powyżej aplikacji można uznać za zwykłe oprogramowanie komputerowe, a nie za szczególne przypadki zastosowania sztucznej inteligencji — chociaż prowadzi nas to na powrót do uwagi McCarthy'ego, że kiedy coś zaczyna działać, przestaje się to określać mianem sztucznej inteligencji. Rozróżnieniem istotniejszym dla naszych celów jest to pomiędzy systemami o wąskim zakresie zdolności poznawczych (niezależnie od tego, czy opatrzymy je etykietką SI, czy nie) a systemami pozwalającymi na rozwiązywanie problemów ze znacznie szerszego spektrum. Zasadniczo wszystkie obecnie wykorzystywane systemy należą do pierwszego, wąskiego typu, jednakże wiele z nich zawiera składowe, które mogą odegrać rolę w przyszłych systemach bardziej ogólnej sztucznej inteligencji

lub posłużyć do ich rozwoju — komponenty takie jak: klasyfikatory, algorytmy wyszukiwawcze, systemy planowania i rozwiązywania problemów i struktury reprezentacji danych.

Jednym z niezwykle intratnych, a przy tym ekstremalnie konkurencyjnych środowisk, w których działają dziś systemy SI, jest globalny rynek finansowy. Zautomatyzowane systemy przeprowadzania transakcji na giełdach papierów wartościowych są obecnie szeroko wykorzystywane przez główne domy inwestycyjne. O ile niektóre z nich są prostymi metodami automatyzacji wykonania konkretnych zleceń kupna lub sprzedaży wydanych przez menedżera funduszu inwestycyjnego, o tyle inne realizują skomplikowane strategie handlowe, które dostosowują się do zmieniających się warunków rynkowych. Systemy analityczne mają do dyspozycji obszerny zbiór technik eksploracji danych oraz analizę szeregów czasowych, która pozwala im na wyszukiwanie wzorców i trendów na rynkach papierów wartościowych oraz na korelowanie ruchów cen historycznych z zewnętrznymi zmiennymi takimi jak słowa kluczowe na paskach wiadomości wyświetlanych na kanałach telewizji informacyjnych. Dostawcy informacji finansowej sprzedają strumienie wiadomości specjalnie sformatowane pod kątem wykorzystania przez tego rodzaju systemy SI. Inne systemy specjalizują się w wynajdowaniu okazji dokonania transakcji arbitrażowych w ramach rynków lub pomiędzy nimi albo w transakcjach o wysokiej częstotliwości, w których źródłem zysku są minimalne ruchy cen — dochodzi do nich w czasie milisekund (w tej skali czasowej opóźnienia w komunikacji nawet przy przesyłaniu sygnałów z prędkością światła za pośrednictwem kabli światłowodowych stają się do tego stopnia znaczące, że opłaca się umieszczanie komputerów w pobliżu giełd papierów wartościowych). Handel algorytmiczny stanowi już ponad połowę handlu akcjami na rynku amerykańskim[69]; to właśnie on obarczany był częściowo odpowiedzialnością za tzw. Flash Crash z 2010 roku (patrz: ramka 2.).

## Opinie na temat przyszłości sztucznej inteligencji

Postęp na dwóch głównych frontach — umocnienia fundamentu statystyki i teorii informacji w kontekście uczenia maszynowego z jednej strony oraz praktycznego i komercyjnego powodzenia rozmaitych aplikacji specyficznych dla danych dziedzin czy problemów z drugiej — przywrócił

## Ramka 2. Flash Crash z 2010 roku

Szóstego maja 2010 roku po południu amerykańskie rynki papierów wartościowych spadły już o 4% w wyniku niepokojów związanych z europejskim kryzysem kredytowym. O godzinie 14:32 duży sprzedawca (zespół funduszy inwestycyjnych) uruchomił algorytm sprzedaży w celu pozbycia się sporej liczby kontraktów future E-Mini S&P 500, które miały być sprzedawane w tempie określanym wskaźnikiem płynności giełdy aktualizowanym na bieżąco. Kontrakty te zostały zakupione przez algorytmy zaprogramowane do szybkiego eliminowania swoich tymczasowych długich pozycji poprzez sprzedawanie kontraktów innym maklerom. Przy braku popytu ze strony podstawowych kupujących algorytmiczni maklerzy zaczęli sprzedawać E-Minis głównie innym algorytmicznym maklerom, którzy przekazywali je kolejnym algorytmicznym maklerom, tworząc efekt „gorącego kartofla" windujący wolumen sprzedaży, co z kolei przez algorytmy sprzedaży zostało zinterpretowane jako wskaźnik wysokiej płynności, skłaniając je do zwiększenia tempa, w jakim kolejne kontrakty E-Minis wystawiane były na sprzedaż, co jeszcze bardziej przyspieszyło niekontrolowaną wyprzedaż. W pewnym momencie algorytmiczni maklerzy zaczęli się wycofywać z rynku, zmniejszając ogólną płynność przy ciągłym spadku cen. O 14:45 handel E-Minis został zatrzymany przez automatyczny wyłącznik. Kiedy handel został wznowiony (zaledwie pięć sekund później), ceny ustabilizowały się i wkrótce papiery wartościowe zaczęły odrabiać większość strat, ale przez chwilę, na samym dnie kryzysu, bilion dolarów zostało tak po prostu wymazanych z rynku, a efekt domina doprowadził do zawarcia szeregu transakcji dotyczących pojedynczych papierów wartościowych, które to transakcje zostały przeprowadzone po „absurdalnych" cenach, jak jeden cent albo 100 tysięcy dolarów. Po zamknięciu tego dnia rynku przedstawiciele giełd spotkali się z przedstawicielami instytucji nadzoru finansowego i postanowili unieważnić wszystkie transakcje wykonane po cenach o 60% lub więcej odbiegających od poziomu przed kryzysem (uznając takie transakcje za „w oczywisty sposób błędne", a więc możliwe do odwołania *post factum* w ramach istniejących zasad regulujących handel na giełdach papierów wartościowych)[70].

Przywołanie tutaj tego właśnie zdarzenia należy uznać za dygresję, ponieważ programy komputerowe uwikłane we Flash Crash nie były szczególnie inteligentne ani wyrafinowane, a ten rodzaj zagrożenia, jakie stworzyły, różni się zasadniczo od obaw, które podniesiemy w dalszej części tej książki w odniesieniu do perspektyw rozwoju superinteligentnych maszyn. Mimo to te wydarzenia są dobrą ilustracją kilku przydatnych prawd. Jedną z nich jest przypomnienie, że interakcje pomiędzy komponentami, które same w sobie są dość proste (jak w przypadku algorytmów sprzedaży czy algorytmicznych programów

maklerskich umożliwiających handel wysokich częstotliwości), mogą wywołać złożone i nieoczekiwane skutki. Ryzyko systemowe potrafi narastać w układzie w miarę wprowadzania kolejnych elementów i to ryzyko nie jest oczywiste, dopóki sprawy nie przybiorą niekorzystnego obrotu (a czasem nie jest oczywiste nawet wówczas)[71].

Druga prawda polega na tym, że inteligentni profesjonaliści mogą wydać programowi polecenia, opierając się na pozornie zdroworozsądkowych i, wydawałoby się, naturalnych założeniach (na przykład takich, że wolumen sprzedaży jest dobrą miarą płynności rynku), a mimo to ich decyzje mogą wywołać ostatecznie katastrofalne skutki, kiedy to program działa, opierając się na nieugiętej, żelaznej logice poleceń nawet w sytuacji, w której te założenia nieoczekiwanie okazują się już nieaktualne. Algorytm robi to, co ma do zrobienia, i jeśli nie jest to bardzo wyjątkowy rodzaj algorytmu, nie przejmuje się zupełnie tym, że na widok jego absurdalnie niewłaściwych poczynań łapiemy się za głowę, a przerażenie zapiera nam dech w piersiach. Ten temat będziemy jeszcze w tej książce podejmowali.

Trzecia obserwacja w odniesieniu do Flash Crash polega na tym, że o ile automatyzacja przyczyniła się do tego zdarzenia, o tyle przyczyniła się również do jego rozwiązania. Zaprogramowana wcześniej logika automatycznego przerwania transakcji, która zawiesiła handel, gdy ceny kompletnie zwariowały, została ustawiona tak, by wykonała się automatycznie, ponieważ słusznie założono, że wydarzenia mogące wywołać te zjawiska będą się toczyć zbyt szybko, by jakikolwiek człowiek zdążył zareagować na czas. Potrzeba zainstalowania zawczasu wykonywanej automatycznie funkcji bezpieczeństwa — będącej przeciwieństwem nadzoru człowieka sprawowanego w czasie rzeczywistym — znów zapowiada temat, który okaże się istotny podczas naszej dyskusji na temat superinteligentnych maszyn[72].

badaniom nad sztuczną inteligencją część utraconego prestiżu. Może pojawiać się jednak resztkowy efekt kulturowy mający wpływ na społeczność SI, a wynikający z jej wcześniejszej historii, który sprawia, że wielu naukowców głównego nurtu niechętnie przyłącza się do wysiłków uznawanych za nazbyt ambitne. Z tego powodu Nils Nilsson zaliczający się do starej gwardii narzeka, że jego współczesnym kolegom brakuje śmiałości, która napędzała pionierów z jego pokolenia:

Troska o „poważanie" miała, jak sądzę, ogłupiający wpływ na niektórych badaczy SI. Słyszę, jak wygłaszają wypowiedzi w rodzaju „SI była niegdyś krytykowana za swoją krzykliwość. Teraz, gdy dokonaliśmy solidnych postępów, nie powinniśmy ryzykować utraty dobrego imienia". Jednym ze skutków tego konserwatyzmu

było zwiększenie koncentracji na „słabej SI" — tej odmianie SI, która służy wspieraniu ludzkiej myśli — i unikaniu „silnej SI" — tej odmiany, która próbuje zmechanizować ludzką inteligencję[73].

Podobną do Nilssona opinię wyraziło kilku innych pionierów tej dziedziny, między innymi: Marvin Minsky, John McCarthy i Patrick Winston[74].

W ciągu ostatnich kilku lat możemy dostrzec odrodzenie zainteresowania sztuczną inteligencją, które być może przerodzi się we wzmożone wysiłki w kierunku skonstruowania sztucznej inteligencji *ogólnej* (właśnie tej określanej przez Nilssona mianem „silnej SI"). Pomijając już szybszy sprzęt, współczesny projekt skorzysta na ogromnych postępach, które dokonały się w wielu poddziedzinach SI, a bardziej ogólnie: w inżynierii oprogramowania, oraz w dziedzinach pokrewnych, takich jak neurobiologia obliczeniowa. Jednym ze wskaźników popytu na wysokiej jakości informację i edukację jest reakcja na udostępnienie przez Uniwersytet Stanforda jesienią 2011 roku darmowego kursu online wprowadzającego do zagadnień sztucznej inteligencji prowadzonego przez Sebastiana Thruna i Petera Norviga. Zapisało się na niego 160 tysięcy studentów z całego świata (a ukończyły go 23 tysiące)[75].

Opinie ekspertów na temat przyszłości SI są bardzo rozbieżne. Nie ma zgody ani co do skali czasowej, ani co do form, jakie może ostatecznie przybrać sztuczna inteligencja. Prognozy na temat przyszłego rozwoju SI, jak odnotowano w jednym z ostatnich badań, „są równie zdecydowane, co różnorodne"[76].

Chociaż współczesny rozkład przekonań nie został dokładnie zmierzony, możemy wyrobić sobie pewien ogólny pogląd na bazie rozmaitych pomniejszych badań i nieformalnych obserwacji. Bardziej konkretnie, w cyklu niedawno przeprowadzonych badań ankietowano członków kilku związanych z tą tematyką społeczności eksperckich, zadając im pytanie, kiedy — ich zdaniem — dojdzie do wynalezienia myślących maszyn dorównujących inteligencją człowiekowi (HLMI — ang. *human level machine intelligence*), definiowanych jako „systemy, które potrafią wykonywać większość ludzkich zawodów na poziomie przynajmniej równie wysokim jak typowy człowiek"[77]. Wyniki pokazano w tabeli 2. Połączona próbka pozwoliła uzyskać następujące szacunki (mediana): 10% prawdopodobieństwa, że HLMI pojawi się przed rokiem 2022, 50% prawdopodobieństwa, że przed rokiem 2040 i 90% prawdopodobieństwa, że przed rokiem 2075 (respondentów proszono to, by oparli swoje szacunki na założeniu, że „prace naukowo-badawcze nie zostaną przerwane żadnym znaczącym zdarzeniem negatywnym").

**Tabela 2.** Kiedy zostaną wynalezione myślące maszyny dorównujące inteligencją człowiekowi[78]?

| | 10% | 50% | 90% |
|---|---|---|---|
| PT-AI | 2023 | 2048 | 2080 |
| AGI | 2022 | 2040 | 2065 |
| EETN | 2020 | 2050 | 2093 |
| TOP100 | 2024 | 2050 | 2070 |
| Łącznie | 2022 | 2040 | 2075 |

Te liczby należy potraktować z pewnym sceptycyzmem: próbki są raczej niewielkie i niekoniecznie reprezentatywne dla ogólnej populacji ekspertów. Podane liczby pozostają jednak w zgodzie z wynikami innych badań[79].

Wyniki tych badań są również zgodne z pewnymi niedawno opublikowanymi wywiadami z dwudziestoma kilkoma naukowcami prowadzącymi prace w dziedzinach pokrewnych SI. Dla przykładu: długa i owocna kariera Nilsa Nilssona koncentrowała się wokół prac nad problemami wyszukiwania, planowania, reprezentacji wiedzy i robotyki; Nilsson jest również autorem podręczników na temat sztucznej inteligencji, a ostatnio ukończył najbardziej do tej pory kompleksowe i wyczerpujące dzieło na temat historii tej dziedziny[80]. Zapytany o datę możliwego pojawienia się HLMI przedstawił następujące prognozy[81]:

10% szans: 2030
50% szans: 2050
90% szans: 2100.

Sądząc z opublikowanych transkryptów wywiadów, rozkład prawdopodobieństwa profesora Nilssona wydaje się całkiem reprezentatywny dla poglądów wielu specjalistów z tej dziedziny, choć znów należy podkreślić, że wyrażane opinie charakteryzuje duża rozmaitość — są tacy praktycy, którzy myślą znacznie śmielej, z dużą pewnością siebie stwierdzając, że systemy HLMI pojawią się zapewne między rokiem 2020 a 2040, i inni, którzy są równie przekonani o tym, że nigdy do tego nie dojdzie albo że ta perspektywa jest nieskończenie odległa[82]. W dodatku niektórzy ankietowani mają poczucie, że pojęcie „ludzkiego poziomu" sztucznej inteligencji jest niewłaściwie zdefiniowane lub mylące bądź też z innych powodów unikają przedstawienia oficjalnych prognoz ilościowych.

W mojej opinii medianom podanym w wynikach badań eksperckich nie przypisano wystarczająco wysokiego prawdopodobieństwa w odniesieniu

do późniejszych terminów. Prawdopodobieństwo na poziomie 10%, że systemy HLMI nie zostaną opracowane do 2075 ani nawet do 2100 roku (po poczynieniu zastrzeżenia, że „prace naukowo-badawcze nie zostaną przerwane żadnym znaczącym negatywnym zdarzeniem"), wydaje się zbyt niskie.

Historycznie rzecz biorąc, badacze zajmujący się sztuczną inteligencją nie wykazali się szczególnym talentem w zakresie przewidywania tempa postępów we własnej dziedzinie ani formy, którą postępy te mogłyby przybrać. Z jednej strony niektóre zadania, jak gra w szachy, okazały się zupełnie osiągalne, i to dzięki zastosowaniu zaskakująco prostych programów; malkontenci utrzymujący, że maszyny nigdy nie będą zdolne zrobić tego czy tamtego, raz po raz okazywali się być w błędzie. Z drugiej strony bardziej typowym wśród praktyków błędem było niedoszacowanie trudności związanych ze skłonieniem systemu do sprawnego wykonywania codziennych zadań i przeszacowywanie postępów dokonujących się w ich własnym ukochanym projekcie lub technice.

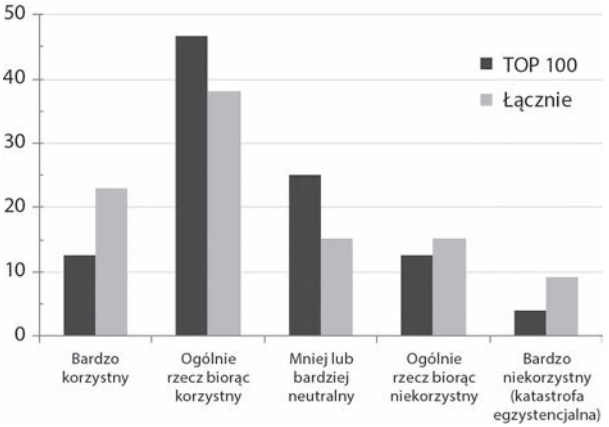
W badaniu postawiono również dwa inne pytania istotne dla naszych rozważań. W pierwszym z nich pytano respondentów, ile czasu, ich zdaniem, zajmie wynalezienie myślących maszyn o nadludzkiej inteligencji, zakładając, że w pierwszej kolejności uda się uzyskać sztuczną inteligencję dorównującą ludzkiej. Wyniki zaprezentowano w tabeli 3.

**Tabela 3.** Ile czasu zajmie przejście od myślących maszyn dorównujących inteligencją człowiekowi do systemów SI o nadludzkiej inteligencji?

| | 2 lata od uzyskania HLMI | 30 lat od uzyskania HLMI |
|---|---|---|
| TOP100 | 5% | 50% |
| Łącznie | 10% | 75% |

W drugim pytaniu poproszono respondentów o odpowiedź na to, jaki długofalowy wpływ na ludzkość będzie miało ich zdaniem wynalezienie systemów dorównujących inteligencją człowiekowi. Odpowiedzi podsumowano na rysunku 2.

I znów moja opinia różni się nieco od opinii wyrażonych w badaniu. Uważam, że bardziej prawdopodobne jest opracowanie systemów superinteligentnych względnie szybko po pojawieniu się HLMI. Mam również bardziej spolaryzowany pogląd na konsekwencje, będąc przekonanym, że opcja „bardzo korzystny" lub „bardzo niekorzystny" jest nieco bardziej prawdopodobna niż wyniki bardziej zrównoważone. Powody, dla których tak sądzę, staną się jasne w dalszej części tej książki.




**Rysunek 2.** Całkowity długofalowy wpływ HLMI na ludzkość[83]

Niewielkie próbki, stronniczy wybór i — przede wszystkim — immanentna zawodność uzyskanych subiektywnych opinii oznaczają, że nie należy zbyt wnikliwie wczytywać się w te badania i wywiady eksperckie. Nasuwają one jednak pewien wątły wniosek. Sugerują, że (przynajmniej wobec braku lepszych danych czy analiz) można rozsądnie przypuszczać, że sztuczna inteligencja dorównująca ludzkiej ma całkiem spore szanse zostać opracowana do połowy tego wieku, a przy tym niezerowe są szanse na to, że pojawi się znacznie szybciej albo znacznie później; że dość duże jest prawdopodobieństwo, że jej stosunkowo rychłym następstwem będą systemy superinteligentne; oraz że znaczące są szanse pojawienia się konsekwencji z bardzo szerokiego spektrum, od bardzo korzystnych do tak niekorzystnych, że równających się zagładzie gatunku ludzkiego[84]. Wszystko to razem sugeruje przynajmniej tyle, że temat wart jest bliższej analizy.

# Skorowidz

## G



## H



## I



## J



## K



## L



## M

## N



## O



## P

## Q



## R



## S

## Ś



## T



## U

## W



## Y



## Z


PROGRAM
PARTNERSKI
GRUPY HELION
1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ
Zmień swoją stronę WWW w działający bankomat!
Dowiedz się więcej i dołącz już dzisiaj!
http://program-partnerski.helion.pl
GRUPA
Helion

## Superinteligencja. Czy cyfrowe umysły zagrażają człowiekowi?

Ludzki mózg to dzieło idealne, najdoskonalszy twór w przyrodzie. Jego dominacją zachwiał komputer Deep Blue, który w 1997 roku pokonał w szachy mistrza świata Garriego Kasparowa. Od tego czasu sztuczna inteligencja jest na ścieżce intensywnego rozwoju. Czy dominacja superinteligencji nad człowiekiem jest realną perspektywą? Czy ta przewaga skończy się kresem naszego gatunku?

Niniejsza książka jest odważną i oryginalną próbą znalezienia odpowiedzi na te pytania. Możliwe kierunki rozwoju technologii cyfrowej i jego konsekwencje są trudne do przewidzenia, a rozstrzygnięcie związanych z tym kwestii jest prawdziwym intelektualnym wyzwaniem. Niemniej z niektórych szans i zagrożeń powinniśmy wszyscy zdawać sobie sprawę. Nick Bostrom, wybitny badacz sztucznej inteligencji i transhumanizmu, stara się jasno i klarownie przedstawić konsekwencje coraz większego udziału maszyn w naszym życiu, opisuje możliwe komplikacje i strategie rozwiązania ewentualnych problemów. Lektura tej książki jest niesamowitą podróżą na krańce istoty człowieczeństwa, wyprawą w przyszłość inteligentnego życia i prawdziwie fascynującym doświadczeniem.

Kluczowe zagadnienia ujęte w książce:

- dotychczasowe osiągnięcia w dziedzinie sztucznej inteligencji
- superinteligencja i jej możliwe ścieżki rozwoju
- zagłada ludzkości a rozwój superinteligencji
- kontrola nad umysłem człowieka w przyszłości

**Nick Bostrom** — szwedzki filozof, profesor Uniwersytetu Oksfordzkiego, kierownik Instytutu Przyszłości Ludzkości działającego w ramach Oxford Martin School. Autor licznych prac dotyczących transhumanizmu (idei zakładającej wykorzystanie osiągnięć nauki i techniki w celu przezwyciężenia ludzkich ograniczeń). W 2009 roku otrzymał za swoją pracę Nagrodę Gannona. Magazyn „Foreign Policy" umieścił go na liście 100 czołowych myślicieli świata.

*Gorąco polecam tę książkę.* Bill Gates

*Cenna lektura. Konsekwencje sprowadzenia na Ziemię drugiego gatunku inteligentnego są na tyle dalekosiężne, że zasługują na głęboki namysł.* The Economist

*Siła argumentów Bostroma nie budzi wątpliwości. (...) To wyzwanie badawcze warte jest podjęcia przez najtęższe umysły matematyczne kolejnego pokolenia. Na szali leży los cywilizacji człowieka.* Clive Cookson, Financial Times

onepress

Księgarnia internetowa:
http://onepress.pl

HELION SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
onepress@onepress.pl

książkiklasybusiness

ebook dostępny na:
ebookpoint

ISBN 978-83-289-0327-2

Helion

Cena: 69,00 zł