PSYCHOLOGIA SZTUCZNEJ INTELIGENCJI

Czym jest sztuczna inteligencja? Jak wpłynie ona na społeczeństwo? Czy jest potężniejsza niż inteligencja ludzka? Psychologia AI bada wszystkie aspekty relacji między psychologią a sztuczną inteligencją, zastanawiając się, jak bardzo sztuczna inteligencja może przypominać ludzi i czy oznacza to, że może posiadać jakąś formę samoświadomości. Analizuje, w jaki sposób systemy sztucznej inteligencji zostały zbudowane na ludzkiej inteligencji oraz podobieństwa między mózgami a komputerami, a także obecne ograniczenia sztucznej inteligencji i sposoby ich przezwyciężenia w przyszłości. Analizuje również interakcje ludzi ze sztuczną inteligencją w życiu codziennym, analizując niektóre zagrożenia etyczne i społeczne, takie jak stronniczość w algorytmach sztucznej inteligencji, oraz konsekwencje dla naszej długoterminowej przyszłości, jeśli sztuczna inteligencja przewyższy ludzi w istotnych aspektach. W miarę jak sztuczna inteligencja wyznacza nowe standardy, Psychologia AI odpowiada na kluczowe pytania dotyczące tego, co tak naprawdę znaczy być człowiekiem i jak sztuczna inteligencja wpłynie na nasze życie pod każdym względem, teraz i w przyszłości.

WSTĘP

Sztuczna inteligencja, w skrócie AI, to „nauka i inżynieria tworzenia inteligentnych maszyn, a zwłaszcza inteligentnych programów komputerowych". Według Johna McCarthy'ego, jednego z twórców tej dziedziny, AI jest „związane z podobnym zadaniem wykorzystania komputerów do zrozumienia ludzkiej inteligencji, ale AI nie musi ograniczać się do metod biologicznie obserwowalnych". Niniejsza książka dotyczy sztucznej inteligencji i jej związku z nauką o ludzkiej inteligencji – psychologią. Jest to złożona, wielokierunkowa relacja. Po pierwsze, jak sugeruje McCarthy, ponieważ psychologia i AI mają podobne przedmioty, obie zajmują się inteligencją. Psychologia bada inteligencję w jej naturalnej formie, obserwowanej u ludzi i zwierząt. AI zajmuje się możliwością stworzenia syntetycznej (stworzonej przez człowieka) inteligencji w artefaktach, zwłaszcza komputerach, ale także w artefaktach, które mogą bardziej przypominać zwierzęta, takich jak roboty. Jak zbadamy, niektórzy psychologowie, znani również jako kognitywiści, wykorzystują metody AI do tworzenia i testowania teorii ludzkiej inteligencji. Badacze zajmujący się sztuczną inteligencją (AI) starają się również zapożyczać idee z psychologii i neuronauki – naukowego rozumienia mózgu i układu nerwowego – aby tworzyć nowe formy AI. Chociaż AI nie ogranicza się do kopiowania naturalnej inteligencji, to kiedy podąża inną ścieżką, czasami nazywamy to sztuczną inteligencją obcą. Ze względu na tę dwustronną wymianę idei i spostrzeżeń, AI i psychologia są ze sobą powiązane od najwcześniejszych dni komputerów. W rzeczywistości idea myślących maszyn jest niemal tak stara jak sama psychologia, począwszy od wczesnych greckich myślicieli. Psychologia jest istotna dla AI również w innych aspektach. Po pierwsze, wraz z pojawieniem się AI, ludzie będą z nimi wchodzić w interakcje, a one zmienią społeczeństwo i sposób, w jaki żyjemy. Zrozumienie, jak AI wpłynie na ludzi i jak ludzie będą się do nich odnosić, w tym do tych fizycznych, takich jak roboty, doprowadziło do powstania nowych dziedzin badań znanych jako interakcja człowiek-maszyna i doświadczenie użytkownika (UI). Metody psychologiczne są również wykorzystywane do zrozumienia, jak ludzie postrzegają AI, na przykład naszego stosunku do AI oraz naszych myśli i obaw dotyczących tego, jak AI może wpłynąć na nas w przyszłości. Oczywiście, sztuczna inteligencja już istnieje, więc niektóre z tych skutków już się pojawiają, ale ludzie myślą i martwią się o przyszłość. Na przykład, czy sztuczna inteligencja może stać się superinteligencją, a jeśli tak, to co się stanie dalej? Ta koncepcja, czasami nazywana osobliwością sztucznej inteligencji (SI), jest przedmiotem gorących debat, zarówno co do jej potencjalnego spełnienia, jak i tego, czy wynik (dla ludzkości) byłby dobry, zły, a nawet bardzo, bardzo zły, do tego stopnia, że ludzie zostaliby zastąpieni przez SI, jak to sobie wyobrażają niektóre filmy science fiction. Wszystkie te aspekty relacji psychologii i SI zostaną zbadane . Sekcja 2 przyjrzy się naturze inteligencji, abyśmy mieli jasny obraz tego, co może oznaczać tworzenie inteligentnych artefaktów. Przyjrzymy się różnym aspektom inteligencji u ludzi i

zastanowimy się, czy inteligencja jest jednym, czy wieloma aspektami. Sekcja 3 przyjrzy się podobieństwom i różnicom między mózgami a komputerami. W końcu, sztuczna inteligencja opiera się na założeniu, że komputery są podobne do mózgów pod wieloma względami, co pozwala obu stronom być inteligentnymi. Zbadamy, w jakim stopniu tak jest. Sekcja 4 przyjrzy się niektórym elementom składowym inteligentnych systemów, zarówno naturalnych, jak i sztucznych, od różnych rodzajów rozumowania i kontroli, poprzez procesy takie jak wyszukiwanie i optymalizacja. Zastanowimy się, czy ludzie myślą inaczej niż maszyny i zbadamy nierozwiązane problemy, w tym rolę emocji w podejmowaniu decyzji. Sekcja 5 bada inspirowane mózgiem sztuczne inteligencje, zwane sieciami neuronowymi, które zostały zaprojektowane tak, aby były szczególnie skuteczne w uczeniu się. Przeanalizujemy, jak rozwijane są najnowsze głębokie sieci neuronowe, aby rozwiązać niektóre z najtrudniejszych problemów w sztucznej inteligencji, w tym szczególnie ludzką zdolność języka. Sekcja 6 omawia sztuczną inteligencję ogólną (AGI), formę AI, która jest równie elastyczna i uniwersalna jak inteligencja ludzka. Przyjrzymy się obecnym ograniczeniom AI w porównaniu z inteligencją ludzką oraz temu, czy AI mogłyby lepiej rozumieć świat, gdyby wyposażono je w robotyczne ciała. Omówimy również, jak rozwój obecnej klasy predykcyjnych AI prowadzi do nowych teorii ludzkiej inteligencji. Wreszcie, w sekcji 7, przyjrzymy się, jak ludzie mogą współistnieć z AI, w tym naszemu podejściu do AI i naszym możliwym relacjom z nią. W tym rozdziale rozważymy również niektóre zagrożenia etyczne i społeczne, takie jak stronniczość i błędy w algorytmach AI, oraz konsekwencje dla naszej długoterminowej przyszłości, ponieważ AI przewyższa inteligencję ludzką w istotnych aspektach. W całym rozdziale rozważane będzie pytanie, jak bardzo AI mogą przypominać ludzi. Jeśli potrafią myśleć tak jak my, czy oznacza to, że mogą posiadać jakąś formę samoświadomości? Jeśli tak, to gdzie to nas zostawia? Czy jesteśmy biologicznym odpowiednikiem bardzo złożonych robotów? Zamiast udzielać pełnej odpowiedzi, postawimy te pytania. Jednakże to pokazuje, że związek psychologii i sztucznej inteligencji dotyka sedna naszej tożsamości jako istot myślących i wskazuje na możliwość głębszego zrozumienia, co znaczy być człowiekiem.

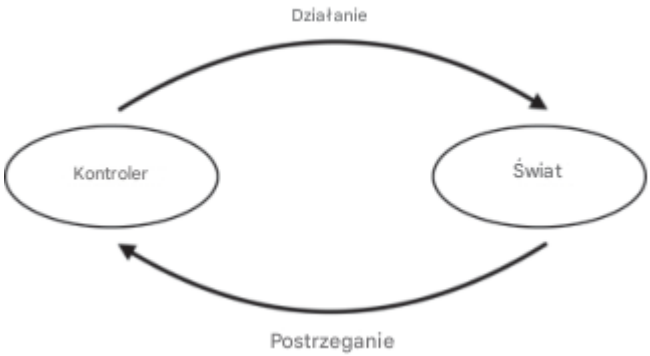
CZYM JEST INTELIGENCJA?

Kiedy się nad tym zastanowić, sztuczną inteligencję można by łatwo nazwać inaczej. Na przykład, można by ją scharakteryzować jako dążenie do stworzenia „sztucznej psychologii" lub zbudowania „syntetycznych umysłów". Umysły są w oczywisty sposób wielowymiarowe, podczas gdy istnieje pogląd, że inteligencja może być tylko jednym z kluczowych elementów i że kiedy to zrozumiemy, wyzwanie AI zostanie rozwiązane. W tym artykule przyjrzymy się początkom badań nad AI i temu, jak doszliśmy do idei tworzenia AI. Przyjrzymy się również powstaniu pojęcia inteligencji w psychologii i temu, czy inteligencja jest jednym, czy wieloma aspektami.

POCZĄTKI SZTUCZNEJ INTELIGENCJI

Inteligentne maszyny wyobrażano sobie od ponad tysiąca lat. Maszyny, które mogły zaskakiwać ludzi pozornie inteligentnym zachowaniem, zaczęły pojawiać się w XVII wieku w postaciautomatów zegarowych. Należała do nich kaczka z miedzi, która zdawała się jeść i trawić, oraz mechaniczny flecista z drewna. Charles Babbage przybliżył ideę sztucznej inteligencji do rzeczywistości w XIX wieku, tworząc pierwszy programowalny komputer cyfrowy, co sprowokowało jego przyjaciółkę Adę Lovelace do sprzeciwienia się możliwości kreatywności maszyn. „Maszyna analityczna [komputer Babbage'a]", pisał Lovelace, „nie ma żadnych pretensji do tego, by cokolwiek zapoczątkować. Może zrobić wszystko, co potrafimy jej zlecić". W pierwszej połowie XX wieku, począwszy od myślicieli takich jak matematyk Norbert Wiener i psychiatra Ross Ashby, znaczna część badań, które dziś moglibyśmy nazwać sztuczną inteligencją, nosiła nazwę cybernetyki. Termin ten pochodzi od greckiego słowa kubernētēs, oznaczającego pełnić rolę pilota lub sternika. Idee cybernetyków kładły nacisk na rolę kontrolera, czyli

mechanizmu decyzyjnego osadzonego w pętli sprzężenia zwrotnego , w ramach którego wyczuwa on świat i generuje odpowiednie działania.



W latach 40. XX wieku Ashby zaprojektował i zademonstrował urządzenie fizyczne zwane homeostatem, które uwzględniało wiele zasad cybernetyki, w tym prawo Ashby'ego o „wymaganej różnorodności" – zgodnie z którym maszyna musi być co najmniej tak złożona, jak system, którym steruje, aby zachować stabilność w obliczu zmian. Cybernetyka przypomina nam, że inteligencja nie występuje w izolacji. Jesteśmy inteligentni w takim stopniu, w jakim działamy właściwie w środowisku, w którym żyjemy, dbamy o siebie i (miejmy nadzieję) osiągamy nasze osobiste cele. Idea inteligencji jako kontroli nie zanikła i zgłębimy ją w dalszej części. Termin „sztuczna inteligencja" nabrał rozpędu dopiero po słynnym spotkaniu badawczym w połowie lat 50. XX wieku. Młody badacz, John McCarthy, który właśnie objął swoje pierwsze stanowisko akademickie, zorganizował ośmiotygodniowe odosobnienie, na które McCarthy, z pomocą kilku czołowych wizjonerów, zdołał zrekrutować jednych z najbardziej aktywnych i znanych badaczy zainteresowanych tworzeniem inteligentnych systemów maszyny. Proponując finansowanie warsztatów, wybrano termin „sztuczna inteligencja", być może częściowo po to, by promować nowość tego przedsięwzięcia w porównaniu z istniejącą dziedziną cybernetyki:

Proponujemy przeprowadzenie dwumiesięcznego, 10-osobowego badania nad sztuczną inteligencją latem 1956 roku w Dartmouth College w Hanover w stanie New Hampshire. Badanie ma się opierać na założeniu, że każdy aspekt uczenia się lub jakakolwiek inna cecha inteligencji może być w zasadzie tak precyzyjnie opisana, że możliwe będzie stworzenie maszyny, która będzie go symulować.

Ostatecznie w warsztatach wzięło udział ponad dwadzieścia osób, a może nawet czterdzieści, choć niektórzy z nich spędzili tam tylko kilka dni. Przez większość dni ta mieszanka inżynierów, matematyków, cybernetyków i psychologów spotykała się na najwyższym piętrze budynku matematycznego Dartmouth i prowadziła ożywione dyskusje na różne tematy. Rozważane tematy obejmowały teorię informatyki, język naturalny, sieci neuronowe, kreatywność i abstrakcję – wszystkie z nich nadal stanowią kluczowe tematy badawcze w dziedzinie sztucznej inteligencji, a niektóre z nich zostaną omówione w dalszej części tej książki. Po rozejściu się uczestników wielu z nich otrzymało znaczne fundusze na badania, aby rozwijać nową dziedzinę sztucznej inteligencji. Niektórzy wierzyli, że stworzenie maszyn dorównujących ludzkiej inteligencji będzie możliwe jeszcze za ich życia. Na przykład w 1960 roku Herb Simon, kluczowy uczestnik warsztatów Dartmouth, czołowy teoretyk sztucznej inteligencji i laureat Nagrody Nobla w dziedzinie ekonomii, napisał, że „maszyny będą zdolne w ciągu dwudziestu lat wykonywać każdą pracę, jaką wykonuje człowiek". W wywiadzie dla magazynu „Life" w 1970 roku Marvin Minsky, jeden ze współorganizatorów zjazdu, zasugerował, że „za trzy do ośmiu lat będziemy mieli maszynę o ogólnej inteligencji przeciętnego człowieka… w tym momencie będzie ona w stanie uczyć się z fantastyczną prędkością" i, co bardziej złowieszcze, że „jeśli będziemy mieli

szczęście, one [AI] mogą zdecydować się na trzymanie nas jako zwierząt domowych". Jak zobaczymy, te przewidywania nie do końca się sprawdziły, choć w latach dwudziestych XXI wieku termin „sztuczna inteligencja" jest bardziej popularny niż kiedykolwiek i (ponownie) nie brakuje wybitnych badaczy przewidujących, że AI wkrótce przewyższą ludzi pod każdym względem. Tak więc utknęliśmy, na dobre czy na złe, z AI jako etykietą wysiłków zmierzających do stworzenia maszyn, które myślą. Jak widać z definicji McCarthy'ego na początku , AI to nauka i inżynieria budowy inteligentnych maszyn. Nasuwa się jednak pytanie, co rozumiemy przez inteligencję? Częściowo po to, by ominąć to wyzwanie, sztuczną inteligencję często definiuje się w odniesieniu do inteligencji ludzkiej; to znaczy, że cokolwiek my, ludzie, robimy, co jest postrzegane jako wymagające inteligencji. Oczywiście, to również nie wyjaśnia inteligencji; raczej zrzuca odpowiedzialność na naukę o ludzkim zachowaniu – psychologię.

PORÓWNANIE LUDZI ZE SZTUCZNĄ INTELIGENCJĄ

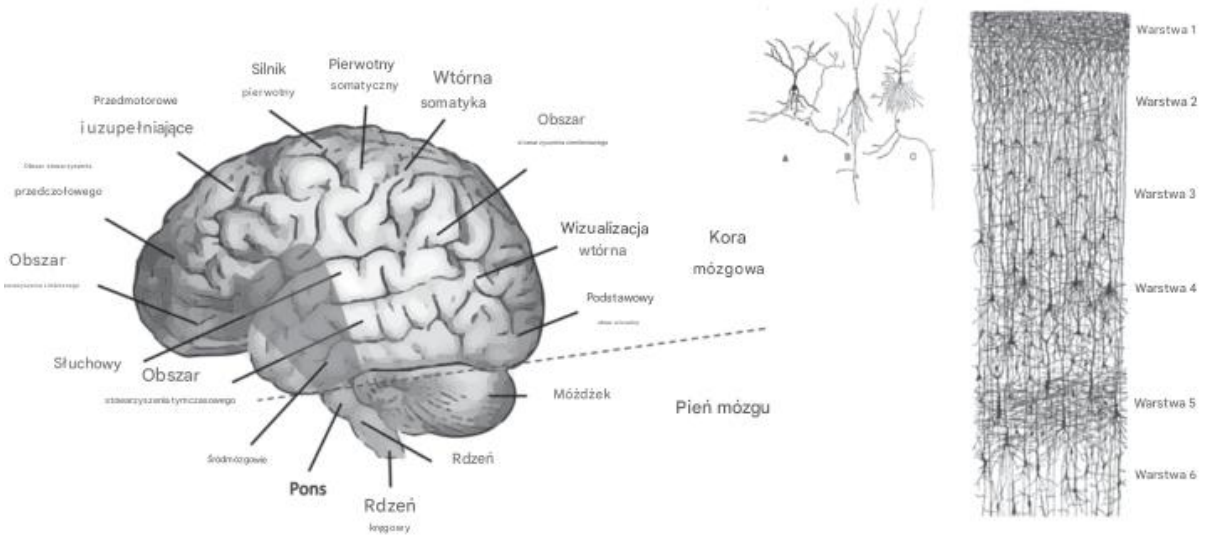
To porównanie z ludźmi sugeruje, że niezależnie od tego, czym jest inteligencja, rozpoznajemy ją, gdy ją widzimy. Ludzie to Homo sapiens – mądry człowiekowaty – wyróżniający się dużym mózgiem, zdolnością do posługiwania się narzędziami i językiem oraz wynalezieniem kultury technologicznej. Chociaż inne zwierzęta są inteligentne, czasami w inny sposób, ludzka inteligencja zawsze była postrzegana jako kluczowy punkt odniesienia dla SI. Ta idea skłoniła Alana Turinga, jednego z twórców informatyki, do zaproponowania testu sprawdzającego, czy maszyny potrafią myśleć, pierwotnie nazwanego „grą imitacji", ale obecnie powszechnie znanego jako test Turinga. Dokładniej, Turing zaproponował, że możemy ocenić, czy maszyna jest rzeczywiście inteligentna, porównując ją z człowiekiem w sesji pytań i odpowiedzi. W jego grze każde pytanie na dowolny temat jest zadawane, wyłącznie w formie pisemnej, zarówno SI (zazwyczaj chatbotowi), jak i człowiekowi. Druga osoba, zazwyczaj osoba posiadająca odpowiednią wiedzę i doświadczenie, jest proszona o ocenę, która odpowiedź pochodzi od człowieka, a która od artefaktu. Jeśli sukces sędziego nie jest lepszy niż zgadywanie (50% trafnych odpowiedzi), to według Turinga powinniśmy się zgodzić, że maszyna jest inteligentna lub „potrafi myśleć", by użyć precyzyjnego sformułowania Turinga. Turing zaproponował swój test w 1950 roku. Od tego czasu był on tematem wielu artykułów, a liczne konkursy zostały zorganizowane, aby zmierzyć sukces SI w tej grze. Jednym z takich konkursów była Nagroda Loebnera, przyznawana od 1990 do 2019 roku, ale nigdy nie przyznano nagrody głównej w wysokości 100 000 funtów. Nagrodę tę mogła wygrać tylko SI, która „zdała" test, zgodnie z oceną jury złożonego z wybitnych sędziów. Filozof John Searle, w słynnej serii artykułów opisujących eksperyment myślowy „Chińskiego Pokoju", argumentował, że nawet gdyby SI zdała test Turinga, nadal nie byłby to dowód na inteligencję maszyny, ponieważ maszyna mogłaby zdać test bez rzeczywistego zrozumienia. Argument Searle'a był niezwykle wpływowy, dlatego ważne jest, aby go tutaj przytoczyć. Searle, jako osoba nieznająca języka chińskiego, wyobraża sobie, że znajduje się w specjalnym pomieszczeniu, którego celem jest odpowiadanie na pytania po chińsku. Osoby mówiące po chińsku poza pomieszczeniem zadają pytania, pisząc je na pocztówkach i umieszczając je pod drzwiami. Aby na nie odpowiedzieć, Searle korzysta z podręcznika zasad napisanego po angielsku, ale opracowanego specjalnie do odpowiadania na pytania po chińsku. Dzięki ścisłemu przestrzeganiu tych zasad, Searle potrafi przetłumaczyć pytania złożone z chińskich znaków – dla niego ciągi bezsensownych „zawijasów" – na odpowiedzi, które również są po chińsku. Udziela odpowiedzi, umieszczając je z powrotem pod drzwiami. Searle wyobraża sobie, że zasady, którymi się kieruje, są na tyle wszechstronne, że dla zewnętrznych obserwatorów jego odpowiedzi mogą być uznane za odpowiedzi rodzimego użytkownika języka chińskiego. Innymi słowy, „Chiński Pokój" może zdać tę odmianę słynnego testu Turinga. Searle utrzymuje jednak, że on (Searle w pokoju) nic nie rozumie z języka chińskiego; po prostu mechanicznie stosuje się do zasad. Podobnie, sztuczna inteligencja, działając zgodnie z wystarczająco zaawansowanym programem, mogłaby zdać test Turinga, ale nie zrozumieć znaczenia generowanych przez siebie odpowiedzi. Dla Searle'a ten prosty eksperyment myślowy był dowodem na to, że test

Turinga nie jest adekwatnym testem inteligencji na poziomie ludzkim i że „programy nie są ani konstytutywne, ani wystarczające dla umysłów". Krytyka „Chińskiego Pokoju" jest niezwykle trafna dzisiaj, gdy współczesne chatboty, takie jak ChatGPT firmy OpenAI, zaskakują nas swoją zdolnością do pisania i prowadzenia konwersacji w sposób bardzo zbliżony do ludzkiego. Trwa spór, czy jest to forma „prawdziwej" inteligencji, czy też ChatGPT i inne współczesne SI, niczym Searle w Chińskim Pokoju, stwarzają pozory inteligencji, jednocześnie nie mając prawdziwego zrozumienia. Ku irytacji niektórych badaczy SI, behawioralne kryteria inteligencji, których przykładem jest zdolność konwersacyjna wymagana do zdania testu Turinga, również wydają się zmieniać z czasem. Na przykład, umiejętność gry w szachy była kiedyś postrzegana jako kluczowy wyznacznik inteligencji na poziomie ludzkim. Jednak komputery pokonują arcymistrzów w szachach od 1996 roku, kiedy komputer szachowy IBM, Deep Blue, pokonał ówczesnego mistrza świata w szachach Garriego Kasparowa w serii sześciu partii. Seria programów komputerowych opracowanych przez Google DeepMind w ostatnich latach triumfowała nad mistrzami w jeszcze trudniejszej grze Go. Jednak sukces w rozgrywkach jest obecnie postrzegany jako mniej istotny wskaźnik niż kiedyś. Obecnie krytycy sztucznej inteligencji częściej wskazują na inteligencję emocjonalną i podkreślają takie zdolności, jak kreatywność, niż biegłość w grach, myśleniu dedukcyjnym, dowodzeniu twierdzeń itd., w których sztuczna inteligencja odniosła znaczący sukces. Można to postrzegać jako „przesuwanie słupków bramkowych" lub jako odkrywanie, jakie są naprawdę trudne aspekty ludzkiej inteligencji. Zatem samo naśladowanie aspektów ludzkiej inteligencji może nie wystarczyć, aby zadowolić niektórych sceptyków wobec myślących maszyn. Czy zamiast tego powinniśmy spróbować usunąć pewne luki w naszych definicjach i bardziej precyzyjnie określić, co oznacza bycie inteligentnym? John McCarthy zaproponował, że inteligencja to „obliczeniowy aspekt zdolności do osiągania celów w świecie". Może to nieco pomóc, przypominając nam, jak podkreślają cybernetycy, że inteligencja nie istnieje w próżni, lecz jest częścią pętli, która pozwala nam działać w świecie. Jednak słowo kluczowe „obliczenia" odgrywa tu dużą rolę. Co to właściwie oznacza? Do tego pytania i do tego, co, jeśli w ogóle, mózgi „obliczają", powrócimy w rozdziale 3. Najpierw jednak przyjrzyjmy się bliżej ludzkiej inteligencji, aby sprawdzić, czy może ona dostarczyć lepszego opisu tego, czego powinniśmy szukać.

KORZENIE POJĘCIA INTELIGENCJI

Współczesne angielskie słowo „intelligence" pochodzi od łacińskiego słowa intelligentia, które z kolei ma dwa rdzenie: inter – „pomiędzy" i legere – „wybierać" lub „czytać". Innymi słowy, oznacza ono coś w rodzaju „wybierać pomiędzy", co przypomina nam o roli inteligencji w określaniu skutecznego działania. Filozof Arystoteles, żyjący w IV wieku p.n.e., jest często uważany za ojca psychologii i osobę, której zawdzięczamy rozróżnienie między percepcją a inteligencją. Dla Arystotelesa aesthesis, czyli percepcja – widzenie, słyszenie, dotykanie, smakowanie itd. – była zdolnością odróżniającą zwierzęta od roślin, ponieważ wszystkie zwierzęta posiadały przynajmniej zmysł dotyku. Natomiast tylko ludzie posiadali nuos – co zwykle tłumaczy się jako inteligencję, intelekt lub umysł. Rozróżnienie Arystotelesa zostało wzmocnione w XVII wieku przez filozofa René Descartesa, słynącego z dualistycznego poglądu na umysł i ciało. Dla dualisty (uwaga, nie kogoś, kto pojedynkuje się!) umysł i ciało to fundamentalnie różne rzeczy, przy czym umysł istnieje poza sferą fizyczną, do której należy ciało. Dla Kartezjusza umysł był unikalny dla ludzi i stanowił siedzibę rozumu, podczas gdy ciało było sferą zmysłów. O ile umysłowi można było zaufać („cogito ergo sum" – myślę, więc istnieję), ciało było potencjalnym źródłem oszustwa. W jednym ze swoich pism Kartezjusz wyobraził sobie złego demona, który manipulował zmysłami, aby dostarczać fałszywych wrażeń ze świata zewnętrznego. Współczesna psychologia stawia pytania o rozróżnienie między inteligencją a percepcją. Oczywiście, obecnie jest oczywiste, że większość zwierząt robi coś więcej niż tylko postrzega i że istnieje o wiele większa ciągłość między inteligencją człowieka a zwierzęcia, niż mogli przypuszczać Arystoteles czy Kartezjusz. Jednakże, co bardziej problematyczne, jeśli przyjrzymy się, jak działają układy nerwowe, rozróżnienie między

percepcją a intelektem również staje się niejasne. Jak zobaczymy w rozdziale 3, układy nerwowe zbudowane są ze specjalistycznych komórek zwanych neuronami. Istnieją neurony czuciowe, które bezpośrednio kodują bodźce ze świata zewnętrznego, takie jak neurony światłoczułe wyścielające wnętrze gałki ocznej, oraz neurony ruchowe, które bezpośrednio stymulują mięśnie, aby wywołać ruch. Pomiędzy komórkami czuciowymi i ruchowymi, wszystkie pozostałe neurony w układzie nerwowym, a więc zdecydowana większość, to interneurony i pełnią nieco inną funkcję. Nawet najprostsze zwierzęta z układem nerwowym, takie jak stułbia – mały organizm słodkowodny o cewkowatym ciele i czułkach, ale bez mózgu – mają sieci nerwowe, które zawierają neurony, które nie są ani czuciowe, ani ruchowe, ale łączą je w złożoną siatkę. Patrząc na mózgi ssaków i ludzi, można określić obszary jako głównie czuciowe lub percepcyjne, a inne jako głównie ruchowe. Jest to jednak bardzo ogólne pojęcie. Obszary czuciowe mogą na przykład brać udział w wyborze działań, a obszary ruchowe mogą odbierać i przetwarzać sygnały sensoryczne. W przypadku obszarów mózgu, które pozornie nie należą do żadnej z tych kategorii, terminologia bywa dość luźna. Na przykład większość kory mózgowej człowieka, składającej się z lewej i prawej kory mózgowej, określana jest jako „asocjacyjna" (patrz rysunek po lewej).



To kolejny termin, który zawdzięczamy Arystotelesowi, który twierdził, że uczenie się polega na kojarzeniu, czyli łączeniu, pojęć mentalnych. Przyglądając się bliżej tym obszarom asocjacyjnym, odkrywamy bardzo szeroką gamę procesów umysłowych, dla których psychologia wypracowała dodatkowe nazwy. Na przykład, uważa się, że istnieją obszary zaangażowane w motywację, emocje, język, rozumowanie, planowanie, nawigację, podejmowanie decyzji itd. Jednak ta etykieta nie jest jednoznacznie ustalona, a wiele obszarów wydaje się przyczyniać do różnych aspektów funkcji, tak jak różne funkcje wydają się być rozproszone w wielu obszarach mózgu. Wydaje się, że mózg nie ewoluował w sposób, który byłby łatwy do interpretacji dla naukowców! Nawet gdybyśmy mogli się zgodzić, że części mózgu można rozróżnić jako głównie percepcyjno-motoryczne lub inne (intelekt), gdy przyjrzymy się tym różnym obszarom mózgu bardziej szczegółowo, wszystkie są zbudowane z podobnych substancji. Na przykład cała ludzka kora mózgowa, czy to sensoryczna, motoryczna, czy asocjacyjna, składa się z sześciowarstwowych sieci neuronowych ułożonych w lokalne struktury zwane kolumnami (patrz rysunek po prawej stronie, i następny rozdział). Neurobiolodzy zgadzają się, że te kolumny mają podobną architekturę sieciową – sposób, w jaki neurony są ze sobą połączone – niezależnie od ich lokalizacji. Chociaż istnieją lokalne różnice, a co najważniejsze, szczegółowe połączenia podlegają korekcie i uczeniu się w trakcie rozwoju, wydaje się, że najważniejsze jest to, jak te fragmenty łączą się ze sobą, a ostatecznie z bodźcami sensorycznymi i bodźcami motorycznymi. Ta

wspólna cecha mechanizmu sugeruje, że problemy percepcji i intelektu mogą nie być tak różne, jak pierwotnie sądzono. Pomimo tych problemów, Arystotelesowskie rozróżnienie między percepcją a intelektem nadal kształtuje sposób, w jaki ludzie powszechnie myślą o umysłach i mózgach. Na przykład, częściej postrzegamy ludzi jako osoby o zróżnicowanych zdolnościach intelektualnych, mimo że istnieje wiele dowodów na istnienie osób o wyjątkowych zdolnościach sensorycznych, takich jak tzw. „superdegustatorzy", którzy potrafią rozróżniać znacznie więcej smaków niż reszta z nas. Podobnie, zdajemy sobie sprawę, że niektórzy ludzie mogą mieć wyjątkowe zdolności sensoryczne i motoryczne, takie jak umiejętność uderzania piłki futbolowej z wyjątkową precyzją. Mimo to jesteśmy mniej skłonni, kulturowo, uznać to za oznakę wyjątkowego intelektu, choć takie umiejętności niewątpliwie angażują znacznie większą część mózgu niż tylko te części, które przetwarzają sygnały z oczu lub kierują ruchem nóg i stóp.

INTELIGENCJA I TESTY ILORAZU IQ

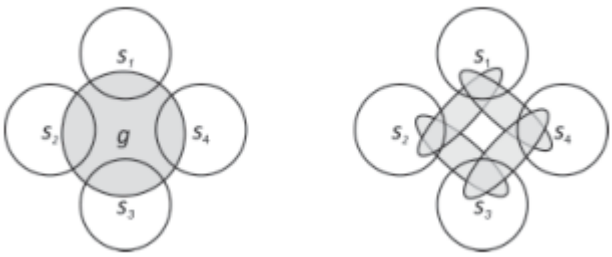
Kluczową ideą psychologii, która często prowadzi nas do myślenia o inteligencji jako o „jednej rzeczy", jest iloraz inteligencji (IQ). Możliwość pomiaru IQ, a tym samym potencjalnego określenia ludzkiej inteligencji, jest jednym z najbardziej znanych osiągnięć dziedziny badań psychologicznych, znanej jako psychologia różnic indywidualnych lub psychologia różnicowa. Jeśli dysponujemy miarą inteligencji, czyż nie oznacza to, że wiemy, czym ona jest? Cóż, nie do końca. Test IQ zawdzięczamy genialnemu psychologowi, Alfredowi Binetowi, działającemu na początku XX wieku. Na zlecenie francuskiego Ministerstwa Edukacji, Binet chciał zidentyfikować dzieci, które słabo radziły sobie w szkole, a zwłaszcza te, które mogłyby skorzystać ze specjalnego nauczania. Pomysł Bineta polegał na zebraniu wielu krótkich zadań lub łamigłówek zaczerpniętych z życia codziennego, które oprócz umiejętności czytania nie wymagałyby żadnej formalnej edukacji. Na przykład zadanie może polegać na zidentyfikowaniu brakującego fragmentu wzoru symetrycznego, znalezieniu brakującej liczby w ciągu lub rozwiązaniu zagadek arytmetycznych, takich jak: „Piotr, który ma dwanaście lat, jest trzy razy starszy od swojego brata. Ile lat będzie miał Piotr, gdy będzie dwa razy starszy od brata?" (Odpowiedź znajdziesz na końcu tego rozdziału, jeśli się zastanawiasz). Podczas gdy wcześniejsze podejścia do pomiaru intelektu koncentrowały się na konkretnych zdolnościach, intencją Bineta było zmierzenie wielu różnych aspektów myślenia i rozumowania w nadziei, że wynik sumaryczny będzie odzwierciedlał ogólne możliwości ucznia. Binet również uporządkował swoje zadania według rosnącego poziomu trudności, oceniając każde z nich według wieku chronologicznego, w którym typowe dziecko powinno być w stanie je wykonać. Prosząc dziecko o przepracowanie sekwencji, aż nie będzie już w stanie poprawnie wykonać zadania, Binet był w stanie oszacować to, co nazwał wiekiem umysłowym dziecka. Odejmując wiek umysłowy od wieku chronologicznego, dawał on miarę tego, jak dziecko radzi sobie w porównaniu z grupą rówieśniczą. Warto zauważyć, że Binet nie zaczął od definicji inteligencji; wręcz przeciwnie, opowiedział się za dużą liczbą i szerokim zakresem zadań, sugerując, że „nie ma znaczenia, jakie testy będą, dopóki będzie ich wiele". To podejście pozwoliło uniknąć problemu,żeby dokładnie określić, co chciał zmierzyć. Binet stworzył trzy wersje swojego testu w ciągu swojego życia. W 1912 roku, wkrótce po śmierci Bineta, William Stern zaproponował podzielenie wieku umysłowego przez wiek chronologiczny (zamiast odejmowania), co doprowadziło do miary, którą obecnie nazywamy ilorazem inteligencji. Narodziło się pojęcie IQ i testu IQ. Testy IQ były i nadal są bardzo szeroko stosowane. Miały jednak również burzliwą historię. Można argumentować, że niewłaściwe stosowanie testów IQ doprowadziło do zmniejszenia szans życiowych wielu osób, a nie do poprawy praktyk edukacyjnych, na którą liczył Binet. Na przykład, kultura testów IQ doprowadziła w niektórych częściach świata do wykluczenia wielu dzieci z dostępu do edukacji i do wymyślania niefortunnych etykiet, takich jak „debil", dla osób o niskim IQ. Chociaż Binet uważał, że mierzy szeroki zakres zdolności intelektualnych i że dzieci z niższymi wynikami mogłyby poprawić jego testy dzięki dodatkowej edukacji, w kręgach psychologicznych w pierwszej połowie XX wieku upowszechnił się zupełnie inny

pogląd. Opierając się na ograniczonych i kontrowersyjnych dowodach, pogląd ten dowodził, że zdolność do dobrego wyniku w testach IQ jest determinowana przez dziedziczność, a nie doświadczenie. Innymi słowy, uważano, że IQ mierzy poziom intelektualny, którego edukacja niewiele może zmienić. Co jeszcze bardziej kontrowersyjne, niektórzy badacze twierdzili, że różnice w IQ mogą wyjaśniać różnice w statusie społecznym – na przykład, że osoby biedne i upośledzone są takie ze względu na niższą inteligencję. Wreszcie, niektóre zakrojone na szeroką skalę badania IQ zdawały się wykazywać różnice w średnim IQ między grupami etnicznymi. Takie ustalenia były wykorzystywane przez frakcje polityczne do uzasadniania nierówności ekonomicznych lub domniemanej wyższości jednej rasy nad innymi. Wiele z tych badań można krytykować za brak uwzględnienia różnic w diecie, wykształceniu i szansach życiowych badanych grup. Czynniki kulturowe w niektórych zadaniach wchodzących w skład testów IQ również sprawiały, że zadania te były łatwiejsze dla niektórych osób niż dla innych. Liczne wady teorii inteligencji opartych na testach IQ zostały omówione w książce Stephena J. Goulda „The Mismeasure of Man". Oprócz tych społecznie dzielących skutków niektórych badań nad IQ, próby pomiaru inteligencji za pomocą testów IQ ugruntowały w kulturze zachodniej przekonanie, że testy IQ mierzą siłę ogólnej zdolności intelektualnej, podobnie jak Arystotelesowskie nuos. Kiedy w 1994 roku Richard Hernstein i Charles Murray opublikowali książkę „The Bell Curve", podsumowali jedną popularną ocenę niemal stu lat badań nad oceną inteligencji: inteligencja to pojedyncza, monolityczna cecha, z którą w dużej mierze się rodzimy, i która różni się w zależności od populacji – większość ludzi znajduje się w pobliżu średniej, a mniejsza liczba osób zbliża się do skrajności (tytułowa „krzywa dzwonowa").

INTELIGENCJE WIELOKROTNE

Ten pogląd na inteligencję jako pojedynczą cechę pozostaje popularny w psychologii akademickiej, podobnie jak w szerszym społeczeństwie, ale istnieją alternatywne szkoły myślenia. W dziedzinie testów inteligencji (IQ) naukowcy szybko opracowali podskale dla różnych aspektów inteligencji. Obecnie, wykonując test IQ, taki jak Skala Inteligencji Dorosłych Weschlera (WAIS), można otrzymać oddzielne wyniki za rozumienie werbalne, pamięć roboczą, organizację percepcyjną i szybkość przetwarzania, a także za ogólne IQ. Chociaż testy tego rodzaju mogą wykazać różnice między poszczególnymi osobami w różnych podskalach, zawsze można wyciągnąć średnią – swoje IQ. Ponadto wyniki w podskalach silnie korelują ze sobą nawzajem i z ogólnym wynikiem; to znaczy, osoby z wysokimi wynikami w jednej podskali mają średnio wysokie wyniki w innych. To samo w sobie nie dowodzi, że istnieje coś, co powinniśmy nazywać inteligencją, co wkrótce zbadamy. David Weschler, pisząc o opracowanym przez siebie teście WAIS, opisał inteligencję jako „zbiorczą lub globalną zdolność jednostki do celowego działania, racjonalnego myślenia i skutecznego radzenia sobie z otoczeniem". Definicja ta pozostawia otwartą kwestię, że inteligencja to coś więcej niż jedna rzecz („zbiór"). W XX wieku kilku wybitnych psychologów zgłębiało koncepcję inteligencji wielorakiej. Na przykład Robert Sternberg zaproponował triarchiczną teorię inteligencji. Jeden rodzaj inteligencji, inteligencja analityczna, odpowiada zdolności rozwiązywania problemów mierzonej w tradycyjnych testach inteligencji. Sternberg argumentował jednak również za inteligencją twórczą, czyli umiejętnością radzenia sobie z nowymi sytuacjami w skuteczny i odpowiedni sposób, oraz za inteligencją praktyczną, czyli zdolnością do radzenia sobie z konkretnymi, realnymi wyzwaniami. Kognitywista Howard Gardner również zaproponował koncepcję inteligencji wielorakiej, zauważając, że ludzie mają różne talenty i że bycie dobrym w jednej dziedzinie nie gwarantuje bycia dobrym w innej. Lista inteligencji Gardnera obejmuje inteligencję językową, logiczno-matematyczną i przestrzenną – ponownie zdolności podobne do tych mierzonych tradycyjnymi testami inteligencji. Do nich Gardner dodał inteligencje bardziej związane ze sztukami kreatywnymi; obejmowały one inteligencję muzyczną i cielesno-kinestetyczną (która może być wykorzystywana w sporcie, tańcu czy aktorstwie). Wreszcie, przynajmniej na oryginalnej liście Gardnera, uwzględnił on inteligencję intrapersonalną, czyli zdolność rozumienia i

efektywnego planowania własnego życia, oraz inteligencję interpersonalną, czyli zdolność rozumienia innych, w tym rozumowania na temat ich uczuć, celów i intencji. Idea odrębnej zdolności do inteligencji interpersonalnej lub społecznej ma długą tradycję. Na przykład, pisząc w 1920 roku, Edward Thorndike, psycholog najbardziej znany ze swojego wkładu w zrozumienie procesu uczenia się, zaproponował, że inteligencja ma odrębne komponenty mechaniczne, abstrakcyjne i społeczne. Inną formą inteligencji pomijaną w tradycyjnych metodach pomiaru inteligencji jest inteligencja emocjonalna. Jak podkreślono w wpływowym artykule Petera Saloveya i Johna Mayera z 1990 roku, odnosi się ona do naszej zdolności do monitorowania własnych uczuć i wykorzystywania ich w sposób dyskryminacyjny do kierowania procesem decyzyjnym. Znaczenie zwracania uwagi na nasze uczucia w podejmowaniu mądrych decyzji zostało również zbadane przez neuropsychologa Antonio Damasio, który przyjrzał się złym decyzjom życiowym podejmowanym czasami przez osoby z uszkodzeniami emocjonalnych części mózgu. Dziennikarz naukowy Daniel Goleman spopularyzował tę ideę w swojej bestsellerowej książce, argumentując, że inteligencja emocjonalna, w tym takie zdolności jak samoregulacja i empatia, ma kluczowe znaczenie dla sukcesu w życiu. Podobny kontrast, zaproponowany przez psychologa Daniela Kahnemana, znanego z prac nad podejmowaniem decyzji ekonomicznych, dotyczy dwóch systemów leżących u podstaw „szybkiego" i „wolnego" myślenia. System szybki jest automatyczny, emocjonalny, często instynktowny i nieświadomy. System wolny jest celowy, świadomy, wymagający wysiłku i logiki oraz rozumowania. Jak mamy rozumieć te wszystkie różne rodzaje inteligencji? Istnieje kilka możliwości. Jedna z nich to istnienie jednej, podstawowej, ogólnej zdolności, obok kilku specyficznych zdolności związanych z tymi różnymi aspektami inteligencji. Idąc za Charlesem Spearmanem29 – jednym z pierwszych, którzy badali te zależności statystycznie – psychologowie często nazywają je „g" i „s". Zgodnie z tą teorią, każda konkretna zdolność, „s", wynika częściowo z ogólnej inteligencji, „g", a częściowo z procesów unikalnych dla tej inteligencji. Tę propozycję ilustruje lewy diagram Venna na rysunku.



Psychologowie, tacy jak Spearman, wywnioskowali, że podobieństwo wyników w różnych podtestach IQ jest spowodowane obecnością ukrytego procesu przyczynowego, g, czyli inteligencji ogólnej. Jest to jeden z wyników zgodnych z potężną metodą statystyczną, również wynalezioną przez Spearmana, zwaną analizą czynnikową, która jest od tamtej pory wykorzystywana do poszukiwania wzorców w danych testujących inteligencję. Jednakże prawy diagram Venna na rysunku może również być zgodny z wynikami w różnych podtestach, które są podobne. W tym przykładzie zilustrowałem ideę, że może istnieć nie tylko jeden ukryty proces, ale kilka. Każda konkretna inteligencja jest w tym przykładzie uwarunkowana przez dwa z tych procesów, ale żaden proces nie wpływa na wszystkie. Alternatywny model inteligencji, opracowany przez Godfreya Thomsona, współczesnego Spearmanowi, sugerował, że może istnieć wiele ukrytych procesów wpływających na różne konkretne miary inteligencji. Analizy przeprowadzone w XXI wieku wykazały, że współczesnych wersji teorii Thomsona nie da się odróżnić statystycznie od teorii g; innymi słowy, obie są równie zgodne z danymi.

INTELIGENCJA – WIELOASPEKTYWNA ZAGADKA

Widzieliśmy zatem, że w psychologii istnieje wiele sposobów myślenia o inteligencji, wiele koncepcji na temat tego, czym jest inteligencja i do czego może być przydatna, ale brak szerokiego konsensusu. Jeden pogląd, moim zdaniem zbyt wąski, głosi, że u podstaw istnieje jeden rodzaj inteligencji – ten, który mierzy się testami IQ – i że leży on u podstaw i wspiera wszystko inne. Zgodnie z tym poglądem, jeśli sztuczna inteligencja potrafi uchwycić tę fundamentalną zdolność rozwiązywania problemów, to może ją zastosować do rozwiązania wszelkiego rodzaju wyzwań, które zazwyczaj uznajemy za wymagające ludzkiej inteligencji. Innym poglądem jest to, że testy inteligencji mierzą tylko jeden aspekt ludzkiej inteligencji i że szerzej liczy się nie nasza zdolność do rozwiązywania pewnych rodzajów łamigłówek, ale nasza zdolność do skutecznego działania w realnym świecie w zakresie codziennych umiejętności i zdolności, które pozwalają nam przetrwać i prosperować. Zgodnie z tym poglądem, inteligencja emocjonalna i społeczna może mieć większe znaczenie niż inteligencja logiczno-matematyczna, jeśli chodzi o odnoszenie sukcesów w naszym wysoce złożonym świecie społecznym. Pokrewny pogląd podkreślałby, że różne rodzaje inteligencji leżą u podstaw szerokiego zakresu ludzkich talentów, a wyspecjalizowane inteligencje pozwalają najwybitniejszym przejawiać olśniewające umiejętności. Należą do nich inteligencje leżące u podstaw kreatywności w muzyce i sztuce, a także inteligencja fizyczna widoczna w sporcie, akrobatyce czy tańcu. Te różne rodzaje inteligencji mogą się częściowo pokrywać, będąc złożonymi z wielu różnych zdolności składowych, ale każdy inny odcień inteligencji może stanowić inną mieszankę. Jeśli ten pogląd na wielorakie inteligencje jest poprawny, to sztuczna inteligencja może rozwijać się w różnym tempie w różnych dziedzinach inteligencji, zamiast robić duże postępy z każdym postępem w kierunku maszynowej wersji inteligencji ogólnej. Jak mamy rozróżnić te podejścia, skoro, jak zauważyliśmy, tak bardzo rozbieżne poglądy mogą potencjalnie odpowiadać za wiele dowodów z testów inteligencji? Psychologia różnicowa, którą tu badamy, jest formą nauk behawioralnych, która stara się zrozumieć ludzi, obserwując ich działania, na przykład ich wyniki w testach inteligencji. Jednak psychologia, a szerzej kognitywistyka, ma w zanadrzu inne narzędzia. Jednym z nich jest neuronauka. Jeśli możemy uzyskać jedynie częściowe zrozumienie ludzkiej inteligencji, analizując zachowania, możemy poszerzyć to zrozumienie, badając podłoża, w których powstaje, czyli w naszych ciałach, a zwłaszcza w ośrodkowym układzie nerwowym i mózgu. Innym ważnym podejściem jest psychologia teoretyczna, która stara się zrozumieć, jak umysł działa na poziomie funkcjonalnym. Innymi słowy, stara się budować teorie, które rozkładają umysł na różne części, wyjaśniają, jak każda z nich działa i jak łączą się one, tworząc całość. Psychologia teoretyczna czerpie inspirację zarówno z nauki o zachowaniu człowieka, jak i z neuronauki, i stara się budować na spostrzeżeniach generowanych przez oba źródła. Teorie umysłu są nieuchronnie bardzo złożone. Dlatego często są systematyzowane i testowane za pomocą modelowania komputerowego, tak jak używamy modeli komputerowych do zrozumienia innych złożonych systemów, takich jak pogoda i gospodarka. Możemy oceniać te modele, testując ich zdolność do wyjaśniania i przewidywania ludzkich zachowań oraz ich spójność z naszą wiedzą o funkcjonowaniu mózgu. W dalszych rozdziałach książki przyjrzymy się bliżej niektórym z tych teorii umysłu, zilustrowanym modelami komputerowymi. Na razie możemy się zgodzić, że dowiedzieliśmy się czegoś o naturze inteligencji, analizując zarówno badania nad inteligencją ludzką, jak i wstępne próby jej powielenia w sztucznej inteligencji, ale wciąż pozostaje wiele do odkrycia

DODATEK

W razie gdybyś nadal miał wątpliwości co do wieku Piotra, oto jak uzyskać odpowiedź. Piotr, który ma 12 lat, jest trzy razy starszy od swojego brata, więc jego brat musi mieć 4 lata. Możemy teraz zapisać równanie $12 + x = 2(4 + x)$, aby pokazać, że próbujemy obliczyć wiek, $12 + x$, w którym Piotr będzie dwa razy starszy od swojego brata, 2 razy $(4 + x)$. Wniosek jest taki, że obaj bracia muszą starzeć się w tym samym tempie, więc x, mimo że nie znamy jeszcze jego rzeczywistej wartości, musi być takie samo po obu stronach. W terminologii informatycznej x jest opisane jako zmienna. Teraz możemy rozszerzyć

nasze równanie, aby otrzymać 12 + x = 8 + 2x, przekształcić (2x–x = 12–8) i uprościć, aby otrzymać x = 4. Zatem Piotr będzie miał 12 + 4 = 16 lat, gdy będzie dwa razy starszy od swojego brata, który będzie miał wtedy 8 lat. Jeśli rozwiązałeś tę zagadkę, możliwe, że przeszedłeś w myślach podobny proces. Inną możliwością jest to, że zamiast rozwiązywać to równanie bezpośrednio, przepracowałeś różne możliwe wartości x, zaczynając od 1 i zwiększając o +1 za każdym razem, aż znalazłeś taką, która działała (obie strony równania były zrównoważone). Obie strategie są przykładami myślenia obliczeniowego.

MÓZGI I KOMPUTERY. RZECZY FIZYCZNE, KTÓRE POTRAFIĄ MYŚLEĆ

Kiedy po raz pierwszy zobaczyłem fragment ludzkiego mózgu na zajęciach laboratoryjnych z psychologii w 1982 roku, byłem zdumiony, że ten dziwnie wyglądający kawałek mięsa, na tyle mały, że można go było trzymać w dłoni, był kiedyś częścią struktury zdolnej do generowania myśli. Umysł i mózg to tak różne rzeczy, że wciąż trudno mi uwierzyć, że ten pierwszy musiał powstać w drugim. Dziś znamy raczej małe kawałki fizycznych rzeczy – na przykład smartfony – które potrafią wykonywać ogromną liczbę złożonych obliczeń niezwykle szybko i obsługują funkcje, które wydają się być umysłowe. Na przykład smartfon w kieszeni będzie zawierał aplikację znaną jako „osobisty asystent cyfrowy", czyli PDA, taką jak Asystent Google lub Siri. Możesz rozmawiać ze swoim PDA, a on odpowie na pytania i udzieli porad. Na przykład teraz mogę poprosić Siri o znalezienie lokalnego sklepu spożywczego, a następnie poprowadzić mnie do niego. Zaskakujące jest, jak szybko przyzwyczailiśmy się do urządzeń przenośnych, które potrafią to robić, i łatwo przeoczyć, jak wiele wymaga rozwiązanie tego rodzaju zapytania. Przyjrzyjmy się temu bliżej. Aby ułatwić mi zakupy spożywcze, mój telefon, wspomagany przez serwery komputerowe w chmurze, musi najpierw przetłumaczyć nagranie mojego głosu na tekst – zadanie, które inżynierowie nazywają rozpoznawaniem mowy. Asystent musi następnie przekonwertować ten tekst – ciąg słów w dowolnym języku, którym mówię – na zapytanie zrozumiałe dla maszyny. Oznacza to, że musi przetłumaczyć mój język na bardziej precyzyjny kod komputerowy, który dokładnie określa, co chcę wiedzieć lub co próbuję zrobić. Ogólnie rzecz biorąc, wiąże się to z wyzwaniem zwanym przetwarzaniem języka naturalnego i może wymagać zaawansowanych algorytmów wyszukiwania w celu zidentyfikowania materiału pasującego do mojego zapytania. Jeśli poproszę PDA o znalezienie pobliskiego sklepu i zaprowadzenie mnie tam, PDA i pakiet systemów oprogramowania, którymi może sterować, muszą „wiedzieć" wystarczająco dużo o sklepach spożywczych, jako miejscach sprzedaży detalicznej określonych rodzajów produktów, i „rozumieć" wystarczająco dobrze, że chcę jeden z nich odwiedzić. Słowa „wiedzieć" i „rozumieć" umieściłem w cudzysłowie, ponieważ, jak widzieliśmy w rozdziale 2, niektórzy filozofowie uważają, że maszyny nigdy nie mogą niczego wiedzieć ani rozumieć. Wrócimy do tego , gdzie zbadamy to, co kognitywiści nazywają problemem uziemienia. Na razie potraktujmy to jako kwestię bardziej praktyczną. Aby udzielać użytecznych porad, PDA musi na przykład odróżniać sklepy od domów prywatnych, sklepy spożywcze od księgarni, sklepów z narzędziami itd. Telefon musi następnie mieć dostęp do mapy, zlokalizować mnie i możliwe sklepy spożywcze na tej mapie oraz określić, które sklepy są najbliżej i jaka byłaby odpowiednia i efektywna trasa, w zależności od mojego środka transportu. Wreszcie, urządzenie musi śledzić moją pozycję w danej chwili, w tym moją orientację, i określić, kiedy udzielić wskazówek nawigacyjnych, dostarczając werbalne i wizualne wskazówki w odpowiednim czasie, aby poprowadzić mnie do celu. Część dotycząca języka mówionego będzie obejmować generowanie wyrażeń języka naturalnego i wykonywanie konwersji tekstu na mowę. Niektóre z tych zadań są łatwiejsze niż inne; Na przykład PDA może uzyskać dostęp do internetowej bazy danych, która przechowuje informacje o sklepach i ich ofercie. Telefon może wykorzystać GPS (Global Positioning System) i sygnały bezprzewodowe do określenia mojej lokalizacji i zastosować algorytm wyznaczania trasy, aby wyznaczyć i poprowadzić mnie po niej. Dekodowanie mowy różnych osób mówiących w wielu językach do niedawna uważano za stosunkowo trudne. Jednak dzięki postępowi w technologii sieci neuronowych maszyny są teraz bliskie ludzkiemu poziomowi umiejętności, a pod pewnymi

względami nawet lepsze, ponieważ rozpoznają dziesiątki języków. Zrozumienie języka, czyli zrozumienie znaczenia wypowiedzi innej osoby, stanowi znacznie większe wyzwanie niż samo rozpoznawanie słów. Jednak PDA są coraz bardziej zdolne do rozpoznawania i nadawania sensu bardzo szerokiemu zakresowi zapytań. Ostatnie postępy w wykorzystaniu dużych modeli językowych zapewniły kolejny znaczący postęp w generowaniu zrozumiałego tekstu i wspieraniu interakcji werbalnej zbliżonej do ludzkiej, co omówimy w rozdziale 5. Wspomniałem tu o sklepach, mapach, lokalizacjach i języku. Wszystkie te koncepcje mają dla Ciebie i dla mnie znaczenie. W Twoim telefonie są to wszystkie dane, co w ujęciu bezwzględnym oznacza wzorce liczb, ostatecznie binarnych jedynek i zer. To uniwersalny kod cyfrowy, na którym opierają się (prawie) wszystkie komputery. Informatycy często mówią o symbolach jako o wzorcach liczb, które mogą coś oznaczać lub reprezentować. Na przykład litera „A" jest często reprezentowana przez liczbę 65, którą w kodzie binarnym można zapisać jako 01000001. Jeśli czytasz tę książkę w formie elektronicznej, oznacza to, że program komputerowy przetwarza plik dokumentu i wysyła instrukcje, aby utworzyć wizualną reprezentację litery „A" na ekranie za każdym razem, gdy zobaczy ten wzór bitów w części tekstowej pliku. Manipulowanie liczbami lub symbolami zgodnie z regułami lub instrukcjami to to, co informatycy nazywają obliczeniami i co daje początek wszystkim niezawodnym i użytecznym zachowaniom Twojego inteligentnego urządzenia. Najprostsze obliczenia, jakie może wykonać komputer, obejmują obliczenia arytmetyczne, porównywanie liczb, zapisywanie liczb w pamięci i przenoszenie liczb z jednej lokalizacji pamięci do drugiej. Złożone obliczenia obejmują łączenie milionów, miliardów lub bilionów tych podstawowych operacji zgodnie z precyzyjną procedurą określoną w programie komputerowym. Aby wykonywać obliczenia i obsługiwać wszystkie te różne przydatne funkcje dla swoich użytkowników, smartfony wykorzystują wiele rodzajów sprzętu fizycznego. Należą do nich wyświetlacze, baterie, aparaty fotograficzne, mikrofony, głośniki itd. Co ważne, sercem każdego inteligentnego urządzenia jest jednostka centralna, czyli CPU – „mózg" urządzenia, jeśli wolisz. Mój iPhone 12 zawiera układ Apple A14. A14 ma sześć rdzeni (procesorów), co oznacza, że może uruchamiać sześć różnych programów jednocześnie. Każdy rdzeń zawiera pamięć, układ logiczny i jednostki sterujące. Kluczowym elementem konstrukcyjnym jest tranzystor, maleńkie urządzenie wykonane z materiału półprzewodnikowego, które pozwala mu znajdować się w jednym z dwóch stanów – włączonym lub wyłączonym, jeden lub zero. Procesory takie jak A14 składają się z miliardów tranzystorów; A14 ma ich aż 11,8 miliarda, a wszystko to w urządzeniu mieszczącym się w kieszeni. Dzięki temu telefon może wykonywać obliczenia z zawrotną szybkością bilionów instrukcji na sekundę. Układ taki jak A14 zawiera zestaw instrukcji obejmujący kilkaset podstawowych operacji. Jednak programiści nie tworzą aplikacji, korzystając bezpośrednio z tych instrukcji, lecz korzystają z języków programowania wysokiego poziomu, takich jak JavaScript, Swift i Python, w których programy można określać znacznie zwięźlej. W każdym z tych języków programy są konstruowane hierarchicznie. Oznacza to, że duże problemy rozwiązuje się, dzieląc je na mniejsze fragmenty zwane procedurami lub funkcjami, które następnie są dzielone na jeszcze mniejsze fragmenty i ostatecznie na elementarne instrukcje w danym języku. Krok polegający na uruchomieniu tego programu w telefonie nazywa się kompilacją i polega na pobraniu całego kodu wysokiego poziomu i przetłumaczeniu go na sekwencje instrukcji opisanych przy użyciu wyłącznie zestawu instrukcji procesora. W przypadku programu komputerowego działającego na smartfonie jest to plik wykonywalny lub aplikacja, która znajduje się w pamięci telefonu i uruchamia się po dotknięciu jej ikony. W 1982 roku technologie te, a ogólnie elektronika użytkowa, były w powijakach. W latach 70-tych pojawiły się pierwsze komputery osobiste i kalkulatory elektroniczne. Pod koniec lat osiemdziesiątych komputery programowalne stały się na tyle małe, że można je było trzymać w ręku. Pierwsze publiczne systemy internetowe pojawiły się pod koniec lat 80., bezprzewodowy internet (Wi-Fi) pod koniec lat 90-tych, a usługi transmisji danych dla telefonów mniej więcej na przełomie tysiącleci. Pierwsze komercyjne systemy rozpoznawania mowy pojawiły się pod koniec lat 80-tych, a pierwsze smartfony w połowie lat 90-tych Pierwsza aplikacja do wyszukiwania głosowego została wydana przez

Google dopiero w 2010 roku, co odzwierciedlało stosunkowo powolny rozwój dokładnego rozpoznawania mowy w wielu językach. Systemy GPS do nawigacji satelitarnej stały się dostępne do użytku publicznego w latach 80-tych, chociaż aplikacje nawigacyjne na telefony pojawiły się dopiero w połowie lat 2000. Pierwszy cyfrowy asystent osobisty, Siri firmy Apple, został wprowadzony na rynek w 2011 roku, krótko po wprowadzeniu wyszukiwania głosowego Google. Zatem w 1982 roku, przed erą telefonów komórkowych, musiałbym znaleźć człowieka, który wskazałby mi drogę do najbliższego sklepu spożywczego, zwłaszcza kogoś znającego okolicę i mówiącego w moim języku. Dziś mogę to wszystko zrobić za pomocą smartfona i PDA. Można by pomyśleć, że telefon i serwery, na których działa, to „tylko" komputery uruchamiające programy, ale powinniśmy zdawać sobie sprawę, że ten zestaw możliwości można uznać za formę sztucznej inteligencji. Twój smartfon wykonuje zadania, które do niedawna mogli wykonywać tylko ludzie, z naszymi stosunkowo dużymi biologicznymi mózgami. Załóżmy, że to sztuczna inteligencja – czy to obca sztuczna inteligencja, czy też to, co robi telefon, jest podobne do tego, co dzieje się w ludzkich głowach?

KRÓTKI OPIS MÓZGU

Mózg, jak każdy inny organ w ludzkim ciele, składa się z miliardów małych części w postaci komórek. Jak już zauważyliśmy, komórki mózgu, które są najważniejsze dla generowania zachowań, nazywane są neuronami. Są to wyspecjalizowane formy komórek, które ewoluowały w celu przetwarzania i komunikacji w organizmie, ale pod innymi względami przypominają każdą inną komórkę w ciele. Na przykład mają jądro komórkowe, które przenosi materiał genetyczny, różne organelle wewnętrzne, takie jak te, które przekształcają pokarm w energię, oraz błonę komórkową, która ogranicza przepływ informacji do wewnątrz i na zewnątrz. Przeciętny ludzki mózg ma około stu miliardów neuronów i znacznie więcej komórek nieneuronalnych. Kluczową różnicą między neuronami a innymi komórkami jest to, że neurony emitują tysiące wypustek, niczym korzenie roślin, których używają do komunikacji z innymi neuronami oraz z komórkami czuciowymi i ruchowymi w całym ciele. Dokładniej rzecz ujmując, neurony wysyłają długie nici, zwane dendrytami, które łączą się z innymi neuronami w specjalnych połączeniach zwanych synapsami. Neurony przekazują również własne sygnały, aby komunikować się z innymi, za pomocą innej specjalnej cienkiej rurki zwanej aksonem. Typowa komórka w przodomózgowiu ma od dziesięciu do trzydziestu tysięcy połączeń synaptycznych, więc ludzki mózg, czyli ludzki konektom, jak się go czasami określa, ma biliony połączeń, być może nawet sto bilionów. Neurony działają poprzez przekazywanie sygnałów w seriach krótkich impulsów elektrycznych zwanych impulsami elektrycznymi. Impulsy elektryczne mają charakter „wszystko albo nic"; to znaczy, że neuron albo generuje impulsy, albo jest spokojny. Impulsy biegnące wzdłuż aksonu wpływają na neurony, z którymi łączą się poprzez pośredniczące synapsy, zwiększając lub zmniejszając prawdopodobieństwo aktywacji komórki po drugiej stronie każdej synapsy. Aktywność w synapsie, która zwiększa prawdopodobieństwo aktywacji kolejnego neuronu, nazywa się pobudzającą, a aktywność, która zmniejsza to prawdopodobieństwo, nazywa się hamującą. Poszczególne neurony są zazwyczaj pobudzające lub hamujące, w zależności od rodzaju synapsy, którą tworzą. Dlatego mózgi można postrzegać jako rozległe sieci neuronów, w których wzorce aktywności mogą rozprzestrzeniać się lotem błyskawicy poprzez połączenia pobudzające lub mogą być tłumione i wygaszane przez działanie neuronów hamujących. Wzory wyładowań dużych populacji neuronów mogą się synchronizować, prowadząc do szczytów i dolin aktywności elektrycznej mózgu, które można opisać jako oscylacje falowe. Podczas snu neurony korowe stają się szeroko zsynchronizowane i oscylują z częstotliwością około 1 Hz, czyli jednym cyklem na sekundę. Jednak w mózgu w stanie czuwania oscylacje zachodzą z wieloma wyższymi częstotliwościami, do 100 Hz i więcej. Na przykład oscylacje między 13 a 30 Hz stają się widoczne, gdy aktywnie angażujesz się w rozwiązywanie problemów. Jedna z popularnych teorii głosi, że ta synchronizacja wyładowań neuronowych pomaga stworzyć spójność rozległych i rozproszonych wzorców wyładowań neuronowych, które obserwujemy w mózgach w stanie czuwania.

Wspomniałem już, że neurony albo wyładowują, albo nie. Przypomina to nieco tranzystory elektroniczne, które omawialiśmy wcześniej, które były albo włączone, albo wyłączone – jeden albo zero. To podobieństwo neuronów do urządzeń cyfrowych skłoniło neurobiologa Warrena McCullocha i jego kolegę matematyka, Waltera Pittsa, do wysunięcia w artykule opublikowanym w 1943 roku tezy, że neurony, podobnie jak komputery, mogą wykonywać pewne rodzaje operacji logicznych. Niemniej jednak, błędem byłoby opisywanie neuronów jako po prostu urządzeń cyfrowych. Owszem, neuron albo wyładowuje, albo nie; jednak czasowy przebieg wyładowań wzdłuż aksonu, wpływ neuroprzekaźników na zachowanie komórki i ogólna biologiczna złożoność komórek sugerują, że dzieje się o wiele więcej, z czego niektóre jeszcze nie rozumiemy. Zastanówmy się najpierw nad czasem wyładowań neuronowych. Niektóre neurony wyzwalają regularne i powtarzalne wzorce w odpowiedzi na bodziec, czasami z częstotliwością proporcjonalną do intensywności bodźca. Ten rodzaj zachowania jest stosunkowo łatwy do zrozumienia. Jednak inne neurony wyzwalają serie wielokrotnych impulsów, po których następują przerwy, podczas gdy inne mogą wytwarzać przerywane impulsy, które są nieprzewidywalne, a nawet losowe. Oprócz aktywności elektrycznej przenoszonej przez aksony i dedryty, neurony komunikują się również chemicznie za pośrednictwem związków zwanych neuroprzekaźnikami i neuromodulatorami. W rzeczywistości większość synaps działa poprzez przekształcanie sygnału elektrycznego po presynaptycznej stronie połączenia w przekaźnik chemiczny. Substancja chemiczna następnie przekracza szczelinę na drugą stronę synapsy, gdzie wyzwala dalszą odpowiedź elektryczną. Substancje chemiczne zaangażowane w transmisję synaptyczną nazywane są neuroprzekaźnikami. Neuromodulatory to substancje chemiczne o bardziej rozproszonym działaniu. Mogą być uwalniane przez neurony lub transportowane we krwi jak hormony, a ich działanie może trwać od kilku milisekund do minut. Istnieje wiele neuromodulatorów, w tym kilka, o których być może słyszałeś, takich jak noradrenalina (zwana również noradrenaliną) i oksytocyna. Na przykład noradrenalina, która jest uwalniana, gdy odczuwamy strach lub stres, pomaga przygotować mózg i ciało do intensywnej aktywności, takiej jak walka czy ucieczka. Oksytocyna, która jest uwalniana, gdy ludzie dotykają się nawzajem w delikatny sposób, pomaga w tworzeniu więzi emocjonalnych, takich jak między rodzicem a dzieckiem, i bierze udział w zachowaniach społecznych i regulacji emocji. Niektóre neurochemikalia mogą działać zarówno jako neuroprzekaźniki (uwalniane lokalnie w synapsach), jak i neuromodulatory (dyfundujące przez mózg). Na przykład dopamina odgrywa kluczową rolę w sygnalizowaniu, gdy wydarzyło się coś niespodziewanie dobrego i jest ważna dla uczenia się. Serotonina pełni wiele funkcji i odgrywa ważną rolę w regulacji nastroju i snu. Chociaż jest ważnym neuroprzekaźnikiem, większość serotoniny w organizmie pochodzi z komórek wyścielających jelita i dociera do mózgu za pośrednictwem krwi. To jeden z wielu sposobów, w jaki części ciała, które nie należą do układu nerwowego, wpływają na sposób myślenia. Wiele leków na zaburzenia nastroju, takie jak lęk i depresja, działa w celu zwiększenia poziomu serotoniny w mózgu. Działanie tych substancji chemicznych w mózgu jest bardziej analogowe, czyli stopniowe lub ciągłe, niż cyfrowe (włączone lub wyłączone). Ogólnie rzecz biorąc, powinniśmy postrzegać neurony nie jako urządzenia analogowe lub cyfrowe, ale jako hybrydę łączącą cechy obu. Co więcej, neurony są niewątpliwie bardziej złożone niż tranzystory, ale nadal możemy je postrzegać jako procesory, czyli urządzenia, które mogą przechowywać pamięć i podejmować decyzje w oparciu o swój poprzedni stan i wzorzec danych wejściowych. Należy pamiętać, że procesor niekoniecznie jest urządzeniem cyfrowym. W rzeczywistości technologie takie jak zegarek naręczny i suwak logarytmiczny można postrzegać jako procesory analogowe (lub komputery), które wykorzystują stale zmieniające się sygnały fizyczne do pomiaru czasu lub wykonywania obliczeń. Współczesna dziedzina neuronauki obliczeniowej, założona przez McCullocha i Pittsa, tworzy modele matematyczne pojedynczych neuronów, traktowanych jako złożone procesory – częściowo cyfrowe, częściowo analogowe. W typowym modelu symulowany neuron pobiera sygnały wejściowe z innych komórek modelowych, modyfikuje je zgodnie z pobudzającą lub hamującą naturą interweniujących synaps i generuje sygnał wyjściowy w postaci

sekwencji syntetycznych impulsów. Wszystko to można opisać za pomocą równań matematycznych, które można symulować na komputerze cyfrowym, uwzględniając wpływ neuromodulatorów, które można uwzględnić jako dodatkowe parametry lub definiując dalsze równania przybliżające sposób dyfuzji tych substancji chemicznych w mózgu. Wiele różnych typów neuronów można obecnie symulować z dużą precyzją. Tak jak potrafimy symulować pojedyncze neurony, neurobiolodzy obliczeniowi potrafią również symulować sieci neuronów, a ostatecznie moglibyśmy sobie wyobrazić symulowanie całych mózgów. W 2023 roku superkomputer HPE Frontier będzie mógł wykonywać milion bilionów operacji zmiennoprzecinkowych na sekundę, co w zupełności wystarczy do symulacji stu bilionów połączeń w ludzkim mózgu, gdybyśmy wiedzieli, jak napisać taki program. Opracowywane są również układy obliczeniowe, które są bardziej bezpośrednio inspirowane sposobem działania i interakcji neuronów biologicznych, co prowadzi do bardziej inspirowanych mózgiem form sztucznych obliczeń. Na przykład, układ A14 w moim iPhonie zawiera już 16-rdzeniowy „silnik neuronowy" zoptymalizowany pod kątem zadań wymagających przetwarzania równoległego, takich jak analiza obrazu, detekcja ruchu i rozumienie języka naturalnego. W dalszej części oraz w rozdziale 5 zbadamy, dlaczego tego typu zadania często wymagają obliczeń przypominających mózg. Prototypowy neuromorficzny układ Loihi firmy Intel ma inne właściwości przypominające mózg, w tym możliwość modelowania 130 000 aktywnych neuronów z maksymalnie 1000 synapsami na neuron. Te sztuczne neurony komunikują się asynchronicznie za pomocą sygnałów przypominających impulsy neuronowe.

MÓZGI JAKO SYSTEMY OBLICZENIOWE

Otwartym pytaniem jest, czy symulowany mózg byłby zdolny do myślenia w taki sam sposób, jak mózg biologiczny. Oczywiście, nigdy nie będziemy w stanie zasymulować wszystkich biologicznych szczegółów tego, co dzieje się w neuronach i mózgach, ale jak blisko musimy się zbliżyć, aby stworzyć sztuczny system, który potrafi myśleć? Używając modeli komputerowych, możemy symulować systemy pogodowe z coraz większą dokładnością i precyzją. Nikt jednak nie wyobraża sobie, że symulowana ulewa mogłaby kiedykolwiek sprawić, że zmokniesz. Podobnie symulacja mózgu nigdy nie będzie miała żadnych fizycznych właściwości chemii mózgu. Dlaczego więc ktokolwiek uważa, że mózg różni się od pogody? Podstawową hipotezą współczesnej kognitywistyki jest to, że mózg jest zasadniczo systemem obliczeniowym; to znaczy wykonuje operacje na danych zgodnie z regułami. Rodzaje obliczeń wykonywanych przez mózg przypominają to, co matematycy określiliby jako operacje statystyczne na dużych zbiorach danych, które charakteryzują się znacznym szumem wewnętrznym. Różnią się one zatem znacznie od ściśle logicznych i bezszumowych operacji konwencjonalnych komputerów. To skłoniło niektórych naukowców i filozofów do argumentowania, że mózgi nie są jak komputery. Jednak sztuczna inteligencja coraz częściej opiera się na bardziej zaszumionych algorytmach statystycznych, aby rozwiązywać najtrudniejsze problemy sztucznej inteligencji. Programy obliczeniowe, które to robią, nazywane są sztucznymi sieciami neuronowymi ze względu na ich oczywistą inspirację sieciami neuronów biologicznych, które znajdujemy w mózgach. Przyjrzymy się im szczegółowo w rozdziale 5. Działając na sprzęcie neuromorficznym, te sztuczne inteligencje coraz częściej wyglądają tak, jakby wykonywały obliczenia, które obserwujemy w mózgach. Czy ma znaczenie, że symulowane mózgi nie mają fizycznych neuronów, synaps lub neuromodulatorów? Prawdopodobnie nie, zgodnie z hipotezą, którą kognitywiści badają co najmniej od czasu wynalezienia komputerów, znaną jako funkcjonalizm. Kluczową ideą jest to, że liczy się nie tyle to, z czego zbudowane są mózgi i komputery – niewątpliwie są zbudowane z bardzo różnych materiałów – ile to, co robią, czyli jaka jest ich funkcja. Wczesnym spostrzeżeniem wspierającym funkcjonalizm był Alan Turing4, który wykazał, że każde urządzenie, które potrafi obliczać dane liczbowe, może w zasadzie emulować dowolne inne. Oznacza to, że moglibyśmy budować komputery ze sznurka i laku, gdybyśmy chcieli (oczywiście, nie jest to praktyczne, stąd „w zasadzie"). Oznacza to również, że wystarczająco

wydajny współczesny komputer cyfrowy, taki jak HPE Frontier, ma potencjał, by naśladować biologiczną hybrydę, taką jak mózg.

OD MÓZGÓW DO UMYSŁÓW

Widzieliśmy, że mózgi to sieci komórek biologicznych, które są silnie ze sobą powiązane i wpływają na siebie nawzajem poprzez sygnały elektryczne i chemiczne, ale w jaki sposób ten system daje początek inteligencji? Chociaż pojedyncze neurony są złożonymi komórkami, niewielu ludzi chciałoby twierdzić, że neurony są indywidualnie inteligentne. Zamiast tego, inteligencja pojawia się, gdy mamy wystarczająco dużą sieć neuronów, aby utworzyć mózg – jak powiedział kiedyś Marvin Minsky: „umysł składa się z rzeczy, które są bezmyślne". Co więcej, aby wykazać się inteligencją i w ogóle się skonfigurować, mózg musi być odpowiednio połączony z innymi częściami ciała – oczami, uszami, kończynami i mięśniami – tak, aby mógł odczuwać i działać. Rozważymy formy ucieleśnionej sztucznej inteligencji, takie jak roboty. Na razie skupimy się na samym mózgu, na tym, jak zbudowany jest z różnych rodzajów neuronów i sieci neuronowych oraz jak łączą się one, tworząc funkcjonalny system, który jest zdolny (wraz z ciałem) do działania, myślenia i doświadczania. Problem budowania mózgów z komórek mózgowych to problem architektury. Architektura oznacza w zasadzie to samo, niezależnie od tego, czy mówimy o budynkach, czy o mózgach. Katedra ma określoną strukturę, funkcję i wygląd, ponieważ różne materiały, z których składa się ta budowla, zostały połączone zgodnie z precyzyjnym i starannie opracowanym planem. Mózg nie jest inny, z tym wyjątkiem, że plan, który powstał w wyniku ewolucji, jest nam całkowicie nieznany. Niczym archeolog odkrywający budynek stworzony przez jakąś zaginioną cywilizację, możemy spróbować odkryć ten plan, badając odkrytą strukturę. W przeciwieństwie jednak do budynku, który jest w dużej mierze konstrukcją stacjonarną, gdzie wygląd dobrze wskazuje cel, mózgi to stale zmieniające się, aktywne struktury, a ich wygląd – duże, przypominające puchate kulki bryły miękkiej tkanki – daje niewiele wskazówek co do funkcjonowania. Zważywszy na miliardy części składowych, próba zrozumienia mózgu na poziomie komórkowym będzie niezwykle trudna, a prawdopodobnie wręcz niemożliwa. Zamiast tego powinniśmy myśleć o mózgu w wielu skalach – od neuronów po sieci, a stamtąd do sieci sieci, które łączą różne części mózgu i łączą mózg z ciałem. Zrozumienie funkcjonalnej architektury mózgu to jedno z największych wyzwań stojących przed współczesną nauką, ale wiele już wiemy. W tym miejscu podsumuję kilka kluczowych idei, które będą przydatne w dalszej części tej książki.

NEURONY, SIECI, MAPY I MIKROOBWODY

Zacznijmy od neuronów. Jak widzieliśmy, istnieje więcej niż jeden rodzaj neuronów – w ludzkim mózgu występuje ich kilkaset, a może nawet tysiące. Zastanawiając się nad funkcjami poszczególnych części mózgu, ważne jest, aby zrozumieć, które rodzaje neuronów są w nie zaangażowane i jak indywidualnie i zbiorowo przyczyniają się do ogólnego wzorca aktywności. Neurony mogą sygnalizować na różne sposoby, jak omówiono wcześniej, na przykład poprzez regularne impulsy lub serie impulsów. Mogą one działać poprzez uwalnianie neuromodulatorów lub pobudzanie włókien mięśniowych. W niektórych sytuacjach pojedynczy impuls z jednego neuronu może przenosić ważne informacje. Częściej, aby podjąć decyzję lub wykonać działanie, potrzebny jest wzór impulsów obejmujący tysiące, a nawet miliony neuronów. Na przykład, w tylnej części gałki ocznej znajduje się warstwa sensoryczna zwana siatkówką, którą możemy traktować jako przedłużenie mózgu. Siatkówka ma ponad sześćdziesiąt różnych typów neuronów, które różnią się kształtem, rozmiarem, neurochemią i, co ważne, wrażliwością na różne rodzaje światła. Komórki, które reagują bezpośrednio na światło wpadające do oka, to pręciki i czopki. Pręciki, których jest około 120 milionów, są wrażliwe na jasność, podczas gdy sześć milionów czopków reaguje na różne długości fal światła, a zatem na kolor – zazwyczaj czerwony, żółto-zielony i niebieski. Po stronie wyjściowej siatkówki znajduje się około 1,2 miliona komórek zwojowych siatkówki, co oznacza znaczną kompresję, w stosunku około 10:1, a inne

formy wstępnego przetwarzania zachodzą już między neuronami światłoczułymi a tą warstwą wyjściową. Oprócz jasności i koloru, te komórki wyjściowe sygnalizują inne atrybuty obrazu wizualnego, w tym kontrast, ruch, migotanie i teksturę. Dla porównania, nowoczesny smartfon będzie miał matrycę co najmniej 12 milionów pikseli (czyli około 1/10 rozdzielczości oka) zbudowaną ze specjalnego rodzaju tranzystora światłoczułego. W formacie RAW każdy piksel jest kodowany jako trzy liczby, po jednej dla koloru czerwonego, zielonego i niebieskiego, z 10-16 bitami na liczbę (tj. od tysiąca do sześćdziesięciu sześciu tysięcy poziomów na kolor). Aparaty cyfrowe również stosują kompresję; na przykład popularny format „jpeg" kompresuje obrazy z niewielką utratą szczegółów, również w stosunku około 10:1. Chociaż istnieje wiele różnic między oczami a nowoczesnymi aparatami cyfrowymi, podstawowa zasada kodowania światła w rozległą gamę sygnałów elektrycznych jest wspólna. Czułość siatkówki jest imponująca — pojedynczy foton światła wystarcza do aktywacji pręcika, chociaż zazwyczaj do zarejestrowania odpowiedzi w mózgu potrzebna jest większa liczba fotonów, docierających mniej więcej w tym samym czasie. A co z sieciami? Kiedy patrzymy na mózg pod mikroskopem, nawet przy małym powiększeniu, widzimy, że niektóre jego części to głównie neurony, zwane istotą szarą, a inne to głównie tkanka łączna, zwana istotą białą. Istnieją również przestrzenie wypełnione płynem, zwane komorami. Istnieje wiele zespołów, zwanych jądrami, silnie połączonych sieci neuronów, a także struktury warstwowe, co już poznaliśmy w przypadku sześciowarstwowej struktury kory mózgowej. Wzory połączeń między neuronami w całym mózgu są wysoce zorganizowane. Większość neuronów ma większość połączeń z sąsiednimi komórkami, ale niektóre mają połączenia dalekosiężne, na przykład te w części ruchowej kory, których aksony sięgają do rdzenia kręgowego, gdzie kontrolują precyzyjne ruchy kończyn i palców. Istnieją również ważne prawidłowości, które dają mocne wskazówki dotyczące funkcji. Na przykład neurony w siatkówce są ułożone w dwuwymiarowe arkusze, dzięki czemu sąsiednie komórki wychwytują światło z pobliskich części środowiska. Ten wzór jest zachowywany, gdy sygnały przepływają przez szlaki mózgu wzrokowego. Gęstość komórek fotoreceptywnych zmienia się w siatkówce, przy czym największe stężenie występuje w dołku czuciowym, w centrum pola widzenia, zmniejszając się około pięćdziesięciokrotnie w kierunku obwodowym. Podobnie, obszary skóry są mapowane na sieci komórek dotykowych w mózgu dotykowym, które również zachowują układ przestrzenny, ale zniekształcają rozmiar przestrzenny. Na przykład obszary czuciowe mózgu odpowiadające za bardzo wrażliwe opuszki palców i usta są znacznie bardziej rozległe niż obszary odpowiadające za duże obszary stosunkowo niereagującej skóry, takie jak środkowa część pleców. Wskazuje to, że mózg poświęca znacznie więcej zasobów na zrozumienie tego, co dzieje się w obszarach skóry, których zazwyczaj używamy do eksploracji świata. Ten geometryczny układ mózgu, wykorzystujący mapy zachowujące sąsiedztwo, jest sposobem na redukcję kosztów połączeń i poprawę efektywności energetycznej. Jest to również sposób kodowania informacji o przestrzennym układzie ciała i świata bez konieczności robienia tego jawnie, pozwalając geometrii wykonać część pracy obliczeniowej. W wielu obszarach mózgu obserwujemy coś, co przypomina mikroobwody, które są powielane w kółko miliony razy. Na przykład w korze mózgowej znajdujemy obwody neuronów o przewidywalnych wzorcach organizacji, które łączą się pionowo przez wszystkie sześć warstw – co odpowiada około 3 mm głębokości (tak, jest tak cienka!) – ale w lokalnym obszarze poziomym o średnicy mniejszej niż pół milimetra . W 1957 roku neurobiolog Vernon Mountcastle zaproponował, że te kolumny korowe, z których każda zawiera około dziesięciu tysięcy neuronów, tworzą modułowe procesory, które w podobny sposób przyczyniają się do funkcjonowania, niezależnie od miejsca, w którym się znajdują. Jedna z popularnych teorii głosi, że mikroobwód korowy jest skonfigurowany do przewidywania własnych sygnałów wejściowych. Jak zobaczymy , badania nad sztuczną inteligencją wykazały, że procesory z taką zdolnością mogą być użytecznym elementem budulcowym, niezależnie od tego, czy próbujesz zrozumieć świat poprzez zmysły, generowanie mowy, czy też działasz w świecie poprzez ruch. Inny rodzaj mikroobwodu znajduje się w tylnej części mózgu, w strukturze zwanej móżdżkiem (łac. „mały mózg"). Chociaż

pierwotnie uważano, że odpowiada on za kontrolowanie ruchu, móżdżek, który zawiera ponad połowę wszystkich neuronów w całym mózgu, obecnie wiadomo, że uczestniczy w wielu różnych funkcjach. Obejmują one ruch i równowagę, mowę i język, przetwarzanie sensoryczne i zachowania społeczne. Podobnie jak kora mózgowa, móżdżek wnosi podobny komponent funkcjonalności do wszystkich tych różnych funkcji. W szczególności mikroobwód móżdżkowy wydaje się być przydatny w dostrajaniu innych układów mózgu, aby były bardziej precyzyjne, na przykład pomaga nam artykułować mowę, precyzyjnie określać czas działań i tak dalej.

NAUKA

Kiedy się rodzimy, prawie wszystkie neurony w naszym mózgu są już obecne. Ale coś ważnego dzieje się w pierwszych kilku miesiącach niemowlęctwa i trwa do końca życia – połączenia między neuronami rosną i zmieniają się. W szczególności zmienia się, czasami bardzo szybko, siła przekazu w synapsach. Wiele z tych zmian jest reakcją na doświadczenie i dlatego stanowi formę uczenia się. Kluczowym odkryciem jest to, że różne części mózgu zdają się realizować różne rodzaje uczenia się. Na przykład kora mózgowa uczy się rozpoznawać wzorce w złożonych sygnałach, takich jak ciągły strumień impulsów przekazujących informacje o kolorze oraz obszarach światła i cienia z oczu. W jakiś sposób kora mózgowa uczy się rozumieć te sygnały, aby rozpoznawać obiekty, miejsca, znajome twarze itd. Ponieważ ta część mózgu wydaje się uczyć bez nauczyciela – być może, jak wspomniano wcześniej, poprzez przewidywanie własnych bodźców – ten rodzaj uczenia się nazywa się uczeniem bez nadzoru. Z drugiej strony, móżdżek wydaje się być zaangażowany w inny rodzaj uczenia się, który opiera się na sygnałach treningowych pochodzących z innych części mózgu. Ponieważ ten proces uczenia się obejmuje zewnętrzny sygnał nauczania, nazywa się go nadzorowanym. Trzeci obszar mózgu, zwany jądrami podstawy mózgu, grupa struktur znajdujących się pod korą mózgową, wydaje się być wyspecjalizowany w uczeniu się na podstawie przerywanych sygnałów treningowych, które wskazują, kiedy wydarzyło się coś nieoczekiwanie dobrego lub złego. Zwykle opisuje się to jako formę uczenia się przez wzmacnianie. Sygnał treningowy istnieje, ale jest on niespecyficzny; nie mówi, gdzie zrobiłeś dobrze, a gdzie źle, a jedynie, czy wynik był lepszy, czy gorszy od oczekiwanego. Ten rodzaj sygnału uczenia się może pochodzić ze środowiska, na przykład, gdy znajdziesz jedzenie lub gdy ktoś, na kim ci zależy, bije ci brawo lub robi ci wyrzuty.

ZASADY ARCHITEKTURY MÓZGU

Ogromna liczba różnych struktur i sieci lokalnych w mózgu jest zagadkowa nawet dla doświadczonych neurobiologów. Czy istnieją jakieś ogólne zasady, które mogłyby pomóc nam zrozumieć, co dzieje się w całym mózgu? Tak, tutaj omówimy tylko kilka. Pierwszą zasadą jest modułowość. System funkcjonalny, taki jak mózg, jest modułowy w tym sensie, że można go podzielić na oddzielne podsystemy, które wnoszą różny wkład. Większość systemów zaprojektowanych przez człowieka jest wysoce modułowa. Pomyślmy o samochodzie. Składa się z tysięcy odrębnych części, które są zorganizowane w podsystemy — silnik, skrzynię biegów, chłodzenie, przekładnię itd. Ciało ludzkie jest również wysoce modułowe; na przykład mamy oddzielne układy oddechowe, krążenia, trawienia, wzroku, słuchu i oczywiście układ nerwowy. Układ nerwowy dzieli się na ośrodkowy układ nerwowy — mózg i rdzeń kręgowy — oraz obwodowy układ nerwowy, który łączy się z ciałem. Ale czy mózg ma również podsystemy? Owszem. Natknęliśmy się już na przykład modularności w poszczególnych kolumnach kory mózgowej. Zauważyłem również, że istnieje bardziej globalna specjalizacja. Na przykład obszary sensoryczne, takie jak kora wzrokowa, zlokalizowane są w tylnej części mózgu, podczas gdy obszary bardziej zaangażowane w kontrolę motoryczną skupiają się z przodu. W tych różnych obszarach kory mózgowej widzimy dalsze dowody modularności. Na przykład w częściach wzrokowych kory mózgowej znajdują się obszary dedykowane przetwarzaniu różnych aspektów widzenia, takich jak kolor, położenie, głębia, ruch itd. W korze słuchowej znajdują się podobszary, które

reagują na dźwięki o różnych długościach fal, w tym te wyspecjalizowane w wykrywaniu dźwięków mowy. Natomiast bliżej tych części kory ruchowej, które odpowiadają za twarz, znajdują się obszary wyspecjalizowane w generowaniu języka ludzkiego, takie jak obszar w lewym płacie czołowym zwany obszarem Broki. Uszkodzenie tego obszaru zazwyczaj nie wpływa na rozumienie języka, ale powoduje trudności w tworzeniu płynnej i poprawnej gramatycznie mowy. Modułowość kory mózgowej oszacowano, badając łączność anatomiczną różnych części mózgu – dosłownie, które części łączą się z którymi, oraz łączność funkcjonalną – które części są aktywne w tym samym czasie i wymieniają się informacjami podczas wykonywania określonych zadań. Oba podejścia dowodzą, że kora mózgowa jest modułowa na wiele sposobów i w różnych skalach. Na przykład analizy anatomiczne wykazały, że korę mózgową można podzielić na cztery ogólne podsystemy: wzrok, słuch, czucie ciała i kontrolę motoryczną oraz planowanie i motywację. Inne analizy funkcjonalne zidentyfikowały sieć struktur i ścieżek o domyślnym trybie działania, które są aktywne, gdy mózg nie jest bezpośrednio zaangażowany w żadne konkretne zadanie. Sieć ta może być ważna dla zapamiętywania przeszłości lub myślenia o przyszłości. W innych częściach mózgu modułowość jest jeszcze bardziej widoczna. Wspomniałem już o wyspecjalizowanym układzie móżdżku, który wnosi podobny wkład do wszystkich różnych podsystemów funkcjonalnych, w których funkcjonuje. Inną wysoce wyspecjalizowaną strukturą jest rdzeń przedłużony, przypominający łodygę przedłużenie rdzenia kręgowego, który zawiera lokalne obwody kluczowe dla wielu funkcji fizjologicznych, w tym bicia serca, ciśnienia krwi i oddychania. Kolejnym przykładem jest jądro nadskrzyżowaniowe, znajdujące się u podstawy mózgu. To jądro zapewnia „zegar biologiczny" i odgrywa ważną rolę w generowaniu i utrzymywaniu rytmu dobowego, który leży u podstaw cyklu snu i czuwania. Osoby z uszkodzeniem tej części mózgu, na przykład w wyniku udaru lub guza mózgu, mogą mieć trudności z utrzymaniem 24-godzinnego rytmu dobowego i często doświadczają okresów czujności i senności o nieregularnych porach w ciągu dnia. Należy zauważyć, że modularność mózgu nie jest tak oczywista jak w systemach skonstruowanych przez człowieka. Na przykład obszary motoryczne mogą odgrywać rolę w przetwarzaniu sensorycznym i odwrotnie, nie ma wyraźnego podziału między zmysłami a myśleniem (intelektem). W dowolnym momencie większość mózgu jest w taki czy inny sposób aktywna, co sugeruje wysoce rozproszony system. Jednak różne części i podsystemy wyraźnie wnoszą zupełnie inny wkład w ogólne funkcjonowanie mózgu w danym momencie. Ta równowaga między przetwarzaniem lokalnie modularnym a globalnie rozproszonym jest jedną z kluczowych zagadek, które naukowcy zajmujący się mózgiem zaczęli rozwiązywać w ostatnich dekadach, pokonując historyczne poglądy, które nadmiernie podkreślały wyłącznie przetwarzanie modularne lub rozproszone. Dowody modularności można postrzegać jako potwierdzenie idei wielorakich inteligencji. Drugą zasadą jest warstwowanie. Natknęliśmy się już na sieci warstwowe w korze mózgowej, ale mózg jest również strukturą warstwową w bardziej zgrubnej skali. Dokładniej, być może wiesz, że mózg dzieli się z grubsza na tyłomózgowie, śródmózgowie i przodomózgowie, które obejmuje korę mózgową . W tych trzech różnych warstwach, które występują u wszystkich kręgowców, występują podobieństwa w funkcjach na wszystkich trzech poziomach.8 Oznacza to, że w przypadku typowych czynności, w które angażują się zwierzęta, w tyłomózgowiu, śródmózgowiu i przodomózgowiu znajdują się struktury, które odpowiadają za ich regulację. Zazwyczaj szlaki w tyłomózgowiu są szybkie, opierają się na sygnałach sensorycznych, które są stosunkowo nieprzetworzone i generują natychmiastowe, niekiedy odruchowe reakcje. Te w śródmózgowiu zazwyczaj wymagają większego stopnia analizy sensorycznej, a zachowania, którymi sterują, są często bardziej złożone, ale często są też wrodzone; to znaczy, że w dużej mierze rodzimy się z tymi zdolnościami. Wreszcie, przodomózgowie rozwija się najwolniej, wykonuje najbardziej złożone analizy świata sensorycznego i generuje najbardziej złożone reakcje, często w wyniku uczenia się. Tam, gdzie powiązane, ale uzupełniające się funkcjonalności nakładają się na siebie, taka architektura sprawia, że mózgi są bardzo wytrzymałe – jeśli jeden podsystem ulegnie awarii, często istnieje inny sposób rozwiązania problemu. Jednak systemy wyższego poziomu zazwyczaj polegają na

działaniu systemów niższego poziomu, ale nie odwrotnie. Oznacza to, że nadal można funkcjonować bez części przodomózgowia, ale tyłomózgowie, które, jak widzieliśmy, bierze udział w regulacji wielu podstawowych funkcji ciała, takich jak oddychanie i bicie serca, stanowi niezbędny interfejs i system kontroli dla reszty ciała. Uszkodzenie krytycznych części tyłomózgowia prowadzi do przedwczesnej śmierci. Trzecią zasadą jest to, że mózgi wykorzystują dynamikę atraktorów, ideę wywodzącą się z fizyki złożonych układów dynamicznych.9 Układ złożony to taki, który składa się z wielu oddziałujących na siebie części, a układ dynamiczny to po prostu taki, który zmienia się w czasie. Nasz układ słoneczny jest złożonym układem dynamicznym – planety i słońce są jego częściami, a ich interakcja wynika z siły grawitacji. Nasze mózgi to kolejny taki system, którego miliardy neuronów stanowią jego części, a ich interakcje wynikają z połączeń synaptycznych, neuromodulacji itd. Nauka o złożonych układach dynamicznych, w skrócie nauka o systemach, pokazuje, że złożone układy dynamiczne mają tendencję do powtarzania w nich preferowanych stanów. Na przykład planety w naszym Układzie Słonecznym mają swoje orbity. Te preferowane stany nazywane są „atraktorami", ponieważ układ będzie dążył do takiego wzorca, zaczynając od stanu losowego, i powróci do niego, jeśli zostanie odwrócony. Twój mózg również będzie miał preferowane stany. Oznacza to, że spośród niezliczonych sposobów, w jakie neurony mózgu mogą oddziaływać na siebie, aby tworzyć wzorce aktywności, niektóre wzorce będą zdecydowanie preferowane nad innymi. Dynamika atraktorów pomaga wyjaśnić, w jaki sposób masowo rozproszony system, taki jak mózg, może wytwarzać spójne zachowania we wszystkich swoich częściach i jak taki system może szybko przełączać się z jednego atraktora, czyli wzorca zachowania, na inny, gdy coś się zmienia, na przykład gdy pojawi się nowy bodziec sensoryczny. Ostatnią zasadą jest to, co nazwę scentralizowanym wyborem działania. Chociaż sieci mózgowe mogą być precyzyjnie dostrojone do wykorzystywania dynamiki atraktorów, wydaje się, że istnieją istotne korzyści z posiadania bardziej scentralizowanych form koordynacji, które gwarantują, że mózg nie będzie próbował wykonywać dwóch niekompatybilnych czynności jednocześnie. Wzdłuż linii centralnej mózgu znajduje się grupa struktur, które wydają się być wyspecjalizowane właśnie w tym celu.10 Należą do nich struktury śródmózgowia i przodomózgowia zwane jądrami podstawy, które, jak już zauważyliśmy, są ważnym miejscem uczenia się. Jądra podstawy wydają się mieć weto wobec wszystkich ruchów dowolnych (tj. ruchów, które nie są odruchową reakcją na bodziec). Oznacza to, że zanim jakakolwiek część mózgu będzie mogła nakazać działanie, musi wysłać sygnały do jąder podstawy, a te struktury muszą następnie usunąć zahamowania z układu ruchowego, aby umożliwić ciału działanie. Podobne szlaki działają wewnętrznie, dzięki czemu jądra podstawy mogą również sterować operacjami poznawczymi (takimi jak myślenie) w innych częściach mózgu. Gdy ten system jest poważnie uszkodzony, na przykład w późnym stadium choroby Parkinsona, ludzie mogą utracić zdolność do podejmowania świadomych działań. Osoby cierpiące na tę wyniszczającą chorobę mogą chcieć się poruszać, ale po prostu nie być w stanie zmusić mięśni do reakcji. Chociaż dynamika atraktorów odgrywa istotną rolę w określaniu reakcji mózgu na świat, jednym z powodów, dla których mózg mógł wykształcić te scentralizowane systemy, jest znaczenie podejmowania szybkich i trafnych decyzji dotyczących dalszych działań. Zwierzęta, które były powolne lub niezdecydowane, byłyby w bardzo niekorzystnej sytuacji, na przykład w unikaniu drapieżników, w porównaniu z tymi, które wykształciły szybkie, wyspecjalizowane systemy precyzyjnego wyboru działań. Działaj szybko albo zostaniesz zjedzony! Rola uczenia się w jądrach podstawy oznacza, że wybór działania może się również poprawić dzięki doświadczeniu. Jeśli działanie wybrane przez jądra podstawy prowadzi do nieoczekiwanej nagrody, uczenie się może dostosować wzorzec połączeń synaptycznych, tak aby to samo działanie było bardziej prawdopodobne w kolejnej podobnej sytuacji. I odwrotnie, jeśli wydarzy się coś złego, prawdopodobieństwo podjęcia takiego działania może być mniejsze następnym razem. Ten proces uczenia się, wykorzystujący neuroprzekaźnik dopaminę jako sygnał uczenia się, wydaje się być kluczowym podłożem dla nabywania nowych umiejętności i kształtowania nawyków w ludzkim mózgu.

ZAMIANA CZĘŚCI MÓZGU ELEKTRONIKĄ

Architektura mózgu, jak właśnie opisano, może wydawać się dość odmienna od tego, co znajdujemy we współczesnych komputerach. Gdzie zatem prowadzi nas to do odpowiedzi na pytanie, czy systemy SI przypominają mózg? Zbadamy te kwestie bardziej szczegółowo w kolejnych rozdziałach, gdzie również zobaczymy, że wraz z tworzeniem SI, które mogą działać w otaczającym nas świecie, obserwujemy coraz większą zbieżność między zasadami funkcjonalnymi, które opisałem dla mózgów, a architekturami obliczeniowymi, które okazują się przydatne w SI i robotyce. Mózgi znacznie różnią się od komputerów pod względem konfiguracji i materiałów, z których są wykonane. Niemniej jednak, nic, czego do tej pory nie odkryliśmy na temat mózgów, nie sugeruje, że ich części składowe, a tym samym ostatecznie całe mózgi, nie mogą być zrozumiane i potencjalnie replikowane za pomocą sztucznych systemów obliczeniowych, czyli komputerów. Jednym z powodów, dla których możemy być tego pewni, jest fakt, że już teraz zastępujemy części układu nerwowego sztucznymi urządzeniami zbudowanymi wokół wyspecjalizowanych komputerów bez zewnętrznych interfejsów, zwanych procesorami wbudowanymi. Urządzenia te działają, aby wykonywać zadanie zastępowanych części lub przywracać funkcjonalność uszkodzonych obwodów. Na zakończenie tego rozdziału omówię kilka niedawnych przykładów. Jednym z urządzeń, które jest już powszechnie stosowane u osób z ubytkiem słuchu, jest implant ślimakowy. Ślimak to część ucha wewnętrznego zawierająca wyspecjalizowane włoski czuciowe, które wykrywają dźwięki o różnych długościach. Gdy fale dźwiękowe docierają do tych włosków, uginają się one, tworząc sygnały sensoryczne, które kodują dźwięk jako sygnały elektryczne przekazywane do różnych części mózgu, takich jak kora słuchowa. Sztuczny ślimak omija uszkodzone ucho, wykorzystując mikrofon wraz z miniaturowym procesorem dźwięku i matrycą elektrod, aby wytworzyć dokładnie taki rodzaj sygnałów elektrycznych w nerwie słuchowym, który przywraca słuch. Poczyniono również postępy w zakresie wymiany siatkówki u osób, które utraciły wzrok. Ponownie, pomysł polega na zastąpieniu neuronów czuciowych (komórek pręcikowych, czopkowych i zwojowych) procesorami, tym razem za pomocą maleńkiej kamery i generującej sygnały elektryczne przekazywane bezpośrednio do siatkówki, a stamtąd do mózgu wzrokowego. Chociaż technologia ta jest mniej rozwinięta niż implant ślimakowy, istnieją urządzenia na wczesnym etapie rozwoju, które przywracają częściowo wzrok osobom, które wcześniej były całkowicie niewidome. Pacjentom z chorobą Parkinsona można wszczepić implanty, które zapewniają formę stymulacji tych części jąder podstawy, które nie funkcjonują prawidłowo z powodu choroby. Odpowiednio dostrojona i ukierunkowana, ta stymulacja elektryczna może skorygować obwód, aby był bliższy jego normalnemu stanowi, łagodząc niektóre objawy choroby Parkinsona, takie jak drżenie i trudności z inicjowaniem ruchu. Niedawno zespół badawczy pod kierownictwem neuroinżyniera Henriego Loracha wykazał, że możliwe jest połączenie uszkodzonego rdzenia kręgowego poprzez umieszczenie wbudowanych procesorów po obu stronach pęknięcia, które komunikują się bezprzewodowo przez Bluetooth. Procesor po stronie mózgu, po stronie pęknięcia, koduje polecenia generowane przez neurony ruchowe w sygnały cyfrowe, które są przekazywane do drugiego procesora po stronie ciała, po stronie pęknięcia, gdzie są one ponownie przetwarzane na sygnały neuronowe, które mogą być wykorzystane do aktywacji mięśni. Jeden pacjent, który doznał zmieniającego życie urazu kręgosłupa w wypadku rowerowym, był w stanie ponownie chodzić i wchodzić po schodach. Inne laboratoria wykorzystały technologię implantów korowych, aby umożliwić sparaliżowanym pacjentom sterowanie ramionami i dłońmi robota za pomocą samej myśli. Technologie opracowywane w laboratoriach badawczych idą jeszcze dalej. Na przykład, naukowcy pod kierownictwem neuroinżyniera Simeona Bamforda wykazali, że mikroprocesor zaprogramowany modelem mikroukładu móżdżkowego może działać u znieczulonego szczura, replikując niektóre efekty uczenia się w móżdżku. Chip odbierał sygnały z wejścia móżdżku i przesyłał je z powrotem do wyjścia móżdżku, omijając w ten sposób biologiczny

mikroukład. Trwają prace nad sprawdzeniem, czy możliwe jest opracowanie podobnych technologii omijających mikroukłady w przodomózgowiu. Docelowo technologie te mogłyby zostać wykorzystane do naprawy uszkodzonych mózgów i potencjalnie do ich rozbudowy, na przykład poprzez umożliwienie urządzeniom cyfrowym działania jako forma pamięci zewnętrznej, która jest bezpośrednio połączona z odpowiednimi obszarami mózgu biologicznego. Prace te mogłyby w przyszłości leczyć schorzenia powodujące amnezję (katastrofalne zapominanie), takie jak choroba Alzheimera. Możliwe, że taki implant mógłby działać tak intuicyjnie, że byłby odczuwany jako rdzeń codziennego myślenia. Oznacza to, że pewnego dnia można by uzyskać dostęp do elektronicznych wspomnień za pomocą samej myśli, tak jak do tych zakodowanych biologicznie. Jeśli uda nam się zastąpić jedną część układu nerwowego elektronicznym substytutem, można sobie wyobrazić wymianę wszystkich jego części. Futurysta Hans Moravec zaproponował to jako eksperyment myślowy. Wyobraź sobie stopniową wymianę wszystkich neuronów w mózgu na ich elektroniczne odpowiedniki, być może po jednym podukładzie lub kolumnie korowej na raz. W którym momencie zacząłbyś czuć i myśleć inaczej? W którym momencie przeszedłbyś z inteligencji biologicznej do sztucznej inteligencji (SI)? Czy powstały w ten sposób mózg elektroniczny nadal posiadałby ludzkie doświadczenie i świadomość? Idee, które omówiliśmy w tym rozdziale, sugerują, że możliwe jest stworzenie sztucznych urządzeń myślących jak mózgi. W następnym rozdziale zagłębimy się w niektóre fundamentalne problemy, które SI już potrafi rozwiązywać, oraz w bloki obliczeniowe wykorzystywane przez te algorytmy. Zbadamy również dokładniej, co może nam to powiedzieć o psychologii ludzkiej inteligencji.

PODSTAWY INTELIGENCJI. ROZUMOWANIE DEDUKCYJNE

W 1989 roku, będąc studentem studiów podyplomowych na kursie „zastosowanej sztucznej inteligencji", jedną z pierwszych rzeczy, których się nauczyłem, było manipulowanie ideami za pomocą logiki. Logiczność kojarzy nam się z komputerami i sztuczną inteligencją, ale jest to również podstawowa umiejętność ludzka, praktykowana od czasów antycznych. Arystoteles, którego poznaliśmy już jako twórcę psychologii, był pod wieloma względami pierwszym kognitywistą, zainteresowanym zrozumieniem i formalizacją natury myślenia poprzez odkrywanie jego zasad. Arystoteles rozwinął jedną z najbardziej znanych form rozumowania logicznego – sylogizm. Rozumowanie sylogistyczne pozwala wyjść od stwierdzeń, o których wiemy lub zakładamy, że są prawdziwe, a następnie zidentyfikować wnioski, które w konsekwencji muszą być prawdziwe. To przykład tego, co nazywamy rozumowaniem dedukcyjnym. Arystoteles napisał: „Dedukcja to mowa, w której, po założeniu pewnych rzeczy, z konieczności, z powodu ich istnienia, coś innego od tych domniemanych rezultatów". To twierdzenie jest sformułowane dość gęsto (mówiąc delikatnie); rozłóżmy je na czynniki, rozważając klasyczny przykład:

Wszyscy ludzie są śmiertelni.

Sokrates jest człowiekiem.

Zatem Sokrates jest śmiertelny.

Mamy tu dwie przesłanki: „wszyscy ludzie są śmiertelni" i „Sokrates jest człowiekiem", z których wyprowadzamy jedną konsekwencję – „Sokrates jest śmiertelny". Wniosek ten nie jest zaskakujący, nawet dla kogoś, kto nigdy nie myślał o logice. Oczywiście, Sokrates jest śmiertelny; w końcu jest człowiekiem! Sylogizm pokazuje, jak można przeprowadzić ten wniosek w bardziej formalny sposób – taki, który wyraźniej uwypukla zachodzące w nim procesy. Arystoteles pokazał konkretnie, jak można automatycznie wyprowadzić nową wiedzę („coś innego") na podstawie wiedzy istniejącej („pewnych rzeczy") poprzez (logiczną) „konieczność". Forma sylogizmu oznacza, że można wziąć tę samą strukturę i zastosować ją do innej treści, mając pewność, że zawsze będzie się w stanie wyprowadzić prawidłowe wnioski (przy założeniu prawidłowych przesłanek). Dlaczego reguły logiki są prawdziwe? Cóż, dla

Arystotelesa było to po prostu kwestią ich „bycia". Można również argumentować, że logika, podobnie jak matematyka, musi być prawdziwa, ponieważ okazała się tak użyteczna w rozumieniu świata. Chociaż pojedynczy sylogizm może wydawać się tak prosty, że nie warto się nim przejmować, siła logicznego myślenia ujawnia się, gdy zaczynamy łączyć ze sobą argumenty. W przypadku sylogizmu nazywa się to polisylogizmem. Rozważmy inny klasyczny przykład:

Pada deszcz.

Jeśli wyjdziemy na zewnątrz, gdy pada deszcz, zmokniemy.

Jeśli zmokniemy, zmarzniemy.

W związku z tym, jeśli wyjdziemy na zewnątrz, zmarzniemy.

Można to zapisać ponownie, aby pokazać, że argument ten składa się po prostu z dwóch połączonych ze sobą sylogizmów w taki sposób, że wniosek z pierwszego staje się przesłanką dla drugiego:

Pada deszcz.

Jeśli wyjdziemy na zewnątrz, gdy pada deszcz, zmokniemy.

W związku z tym, jeśli wyjdziemy na zewnątrz, zmokniemy.

Jeśli wyjdziemy na zewnątrz, zmokniemy.

Jeśli zmokniemy, zmarzniemy.

W związku z tym, jeśli wyjdziemy na zewnątrz, zmarzniemy.

Można to postrzegać w ten sposób, że przyjmujemy formę rozumowania opartego na zdrowym rozsądku i wyrażamy je w taki sposób, że wszystkie kroki podejmowane w celu dojścia do wniosku są bardziej jednoznaczne. Ludzie potrafili rozumować dedukcyjnie przed Arystotelesem i nadal to robią, nie słysząc o logice. Oczywiście wiesz, że jeśli wyjdziesz na deszcz, zmokniesz, a jeśli zmokniesz, możesz zmarznąć. Z tego powodu rozsądnie jest nie wychodzić na deszcz, a jeśli już, to nie brać parasola. Jedna z hipotez głosi, że twój umysł, niekoniecznie o tym wiedząc, doszedł do tego wniosku na podstawie szeregu wnioskowań, podobnych do sylogizmów Arystotelesa (istnieją inne możliwości, o których mogłeś się dowiedzieć z doświadczenia). Podobny proces wnioskowania dedukcyjnego został usystematyzowany w postaci systemów eksperckich AI, znanych również jako systemy oparte na wiedzy, które wykorzystują struktury logiczne złożone z reguł „jeśli-to", podobnych do sylogizmów i nazywanych również regułami produkcji. Podobnie jak sylogizmy, produkcje można sekwencjonować, aby generować nowe wnioski. Nazywa się to wnioskowaniem łańcuchowym w przód i polega na ustalaniu nowych faktów na podstawie istniejących, które wraz z istniejącymi faktami generują kolejne reguły itd. Systemy tego typu są szeroko stosowane w takich obszarach jak obsługa klienta, usługi finansowe, inteligentne korepetycje i diagnostyka medyczna, często za pośrednictwem internetowych chatbotów. Możliwe więc, że korzystałeś z tego rodzaju sztucznej inteligencji, nie zdając sobie z tego sprawy. W XIX wieku logikę Arystotelesa rozwijali inni myśliciele, tacy jak filozof Gottlob Frege, który opracował język formalny – logikę predykatów – do reprezentowania idei i rozumowania o nich. Frege wynalazł ideę zmiennych, zastępując stwierdzenia takie jak „Sokrates jest człowiekiem" bardziej ogólną formą „istnieje x takie, że x jest Sokratesem, a x jest człowiekiem". Logika predykatów utorowała drogę językom programowania, których używamy dzisiaj. Mniej więcej w tym samym czasie matematyk George Boole w książce zatytułowanej „Prawa myśli"3 rozwinął to, co dziś nazywamy logiką symboliczną, wykorzystując skrót symboli, takich jak x i y, zamiast wyrażeń języka naturalnego do przeprowadzania rozumowania dedukcyjnego. Zdania można następnie powiązać z wartościami

logicznymi; to znaczy, można je zidentyfikować jako prawdziwe lub fałszywe, a następnie połączyć z operatorami logicznymi, takimi jak i, lub i nie. Wszystko to można oczywiście wyrazić za pomocą liczb. Rzeczywiście, podstawowa dwoistość logiki – prawda i fałsz – idealnie odwzorowuje się na binarne wartości jeden i zero. Dedukcja jest tak potężnym procesem, że leży u podstaw większości zasad i idei, które mamy w matematyce. W istocie, prawie wszystkie ważne idee matematyczne zostały potwierdzone w procesie rozumowania dedukcyjnego, wychodząc od istniejących zasad, które zostały wcześniej wydedukowane. Gottlob Frege jest uznawany za sformalizowanego ideę matematycznego „dowodu". Należy jednak pamiętać, że nie oznacza to, iż wszystkie idee w matematyce powstały w wyniku dedukcji. Raczej, po dojściu do nowego wniosku, matematyk zazwyczaj będzie sięgał do dedukcji, aby udowodnić, że ta nowa idea jest koniecznie poprawna. W istocie, w dobie sztucznej inteligencji matematycy często korzystają z automatycznych dowodzeń twierdzeń, aby pomóc w ustanowieniu formalnego dowodu hipotezy matematycznej. W tej książce zbadamy inne sposoby, w jakie idee mogą powstawać, jednocześnie uznając, że procesy leżące u podstaw prawdziwej oryginalności myśli pozostają nieco tajemnicze! Kolejnym wyzwaniem dla dedukcji jest to, że mając kilka przydatnych informacji, możemy wywnioskować inne; jednak większość tych dedukcji może być dość banalna i raczej bezużyteczna. Na przykład, jeśli wyjdę na zewnątrz podczas deszczu, mogę wnioskować, że zmoknę, że moje ubrania zmokną, że moje włosy zmokną (jeśli nie będę miał na sobie czapki) i tak dalej. To wszystko jest bardzo prawdziwe, ale być może niezbyt pomocne. Jak, w logice dedukcyjnej, możemy priorytetyzować wyprowadzanie użytecznych wniosków istotnych dla naszych celów i potrzeb? Kolejny problem, zwany problemem ramy, został zidentyfikowany przez Johna McCarthy'ego i jego współpracownika Patricka Hayesa.4 Jeśli coś zmienia się w świecie, ile istniejących faktów, które już wywnioskowałem, muszę ponownie ocenić? Jeszcze szerzej, jak mogę zdecydować, o czym warto rozumować, a co nie? To wskazuje, że wszelkie myślenie nie może być jedynie dedukcją.

INDUKCJA I PRZYCZYNOWOŚĆ

Gdyby dedukcja była jedynym sposobem odkrywania czegokolwiek, nasza wiedza byłaby raczej ograniczona. Dedukcja ujawnia rzeczy, które są prawdziwe, ze względu na inne rzeczy, o których już wiemy, że są prawdziwe. W pewnym sensie nigdy nie mówi nam niczego naprawdę nowego, ponieważ, gdybyś był szczególnie pedantyczny, mógłbyś argumentować, że skoro znasz podstawowy zestaw faktów i reguł logiki, to wszystko inne, co możesz wydedukować, jest już implicite znane. Jedną z dziedzin zajmujących się odkrywaniem innego rodzaju wiedzy jest nauka. Rzeczywiście, kiedy przyglądamy się nauce, widzimy, że jej podstawową metodologią nie jest dedukcja, choć z pewnością odgrywa ona pewną rolę, ale to, co możemy nazwać indukcją. Indukcja postępuje poprzez obserwację; patrzymy na świat i widzimy, że pewne wzorce powtarzają się w kółko. Na przykład, przez niezliczone pokolenia nasi przodkowie obserwowali, że słońce wschodzi na wschodzie i zachodzi na zachodzie. Powtarzane obserwacje wschodów i zachodów słońca, cykli ruchu Księżyca, planet i innych ciał niebieskich, prowadzone przez tysiące lat, doprowadziły do powstania pierwszych kalendarzy, a ostatecznie do możliwości przewidywania położenia planet i Słońca z zadziwiającą dokładnością. W XVI wieku doprowadziło to do rewolucji kopernikańskiej – zrozumienia, że planety krążą wokół Słońca – a później pozwoliło Johannesowi Keplerowi na dalsze wywnioskowanie, że orbity planet mają kształt elipsy. Obserwacje astronomiczne odegrały również kluczową rolę w obaleniu poglądu Izaaka Newtona na absolutną przestrzeń i czas oraz zastąpieniu go teorią względnej czasoprzestrzeni Alberta Einsteina. Zasada indukcji jest prosta – powtarzalne obserwacje prawdopodobnie wynikają z jakiejś ukrytej regularności, którą należy odkryć. Mówiąc ogólniej, obserwujemy pewien rozkład wyników i szukamy wyjaśnienia, dlaczego wystąpił właśnie ten, a nie inny rozkład. Ostatecznie nauka dąży do wyjaśnienia, kontrolowania i przewidywania otaczającego nas świata5 w oparciu o dowody zebrane drogą indukcji. Czyni to, próbując znaleźć najprostsze i najskuteczniejsze wyjaśnienie zbioru obserwacji. Obecnie często osiąga się to za pomocą modeli komputerowych, które opisują i ujmują w kategoriach

matematycznych zachowanie interesującego nas układu fizycznego lub biologicznego. Nasze modele Układu Słonecznego i pogody działają w ten sposób, podobnie jak nasze modele mózgu. Coraz częściej algorytmy uczenia się są wykorzystywane do konstruowania nowych reguł wnioskowania i modeli przyczynowych na podstawie danych w dużych systemach sztucznej inteligencji. Takie podejścia były wykorzystywane do rozwiązywania wcześniej trudnych problemów naukowych, takich jak dedukcja kształtu białka na podstawie jego składu chemicznego. Rozumowanie indukcyjne i budowanie modeli są zatem równie ważne, jak rozumowanie dedukcyjne, jako kluczowe elementy składowe sztucznej inteligencji. Bardziej przyziemnym przykładem jest filtr antyspamowy używany przez dostawcę poczty e-mail, który może wykorzystywać procesy indukcyjne do identyfikowania i etykietowania potencjalnych wiadomości spamowych.

WYSZUKIWANIE I OPTYMALIZACJA

Dziś, myśląc o wyszukiwaniu, często mamy na myśli wyszukiwarki takie jak Google, Bing i DuckDuckGo, z których większość z nas korzysta codziennie, aby znaleźć informacje w internecie. Nowoczesne wyszukiwarki same w sobie są złożonymi systemami sztucznej inteligencji (SI), które wykorzystują narzędzia takie jak przetwarzanie języka naturalnego, rozpoznawanie wzorców i analiza zachowań użytkowników, a także zaawansowane algorytmy rankingowe, aby starać się udzielać użytecznych odpowiedzi na nasze zapytania. Jednak wyszukiwanie, szerzej rozumiane, jest jednym z najbardziej fundamentalnych wyzwań w SI – wiele problemów, które SI próbuje rozwiązać, od wyboru ruchów szachowych po lądowanie robota-łazika na Marsie, można scharakteryzować jako problemy wyszukiwania. Aby zastosować wyszukiwanie w SI, najpierw definiuje się przestrzeń stanów, a następnie stosuje procedurę wyszukiwania, aby zbadać tę przestrzeń i znaleźć stany odpowiadające dobrym rozwiązaniom. We wczesnych latach rozwoju SI ludzie pracowali z problemami, w których przestrzeń stanów była łatwa do zdefiniowania, takimi jak tradycyjne gry planszowe. Gry adwersarskie, takie jak szachy, można postrzegać jako wyszukiwanie w grafie . Stan początkowy to pierwszy węzeł na grafie, a każdy możliwy ruch prowadzi do nowego węzła połączonego z węzłem początkowym krawędzią. Każdy nowy węzeł wypuszcza nowe krawędzie i kolejne węzły itd. Graf gry takiej jak szachy bardzo szybko staje się bardzo gęsty – już po pierwszych trzech ruchach w szachach istnieje prawie dziewięć tysięcy możliwych stanów. W 1950 roku Claude Shannon, twórca teorii informacji i współorganizator Dartmouth Summer School on AI, podjął próbę obliczenia liczby możliwych stanów w grze w szachy. Jego szacunek 10^{120} sugeruje, że szachy mają więcej możliwych stanów niż atomów w obserwowalnym wszechświecie. Ponadastronomiczny rozmiar przestrzeni stanów dla problemów tak prostych do opisania jak szachy – trzydzieści dwie figury na siatce osiem na osiem – pokazuje nam, że strategie wyszukiwania dla AI muszą być naprawdę bardzo dobre. Nawet najszybszy komputer szachowy nie może pozwolić sobie na zbadanie wszystkich potencjalnych ścieżek do zakończenia gry podczas wyboru kolejnego ruchu. Na przykład, superkomputer szachowy, który mógłby analizować bilion stanów szachowych na sekundę, potrzebowałby stu bilionów lat na przeanalizowanie zaledwie $3{,}15 \times 10^{32}$ stanów, co stanowi niewielki ułamek szacunków Shannona dotyczących całkowitej liczby możliwych stanów. Według niektórych szacunków, za sto bilionów lat zgaśnie ostatnie światło w naszym rozszerzającym się wszechświecie. Oczywiste jest, że wyczerpujące wyszukiwanie, nawet w dobrze zdefiniowanej przestrzeni problemów, takiej jak szachy, nie jest realną opcją. Gwałtowne rozprzestrzenianie się stanów w wielu problemach wyszukiwania jest często określane jako „eksplozja kombinatoryczna". Aby sprostać temu wyzwaniu i efektywnie przeprowadzać wyszukiwanie w problemach z dużymi przestrzeniami stanów, sztuczne inteligencje wykorzystują algorytmy, które oceniają różne stany pod kątem ich relatywnie lepszej lub gorszej kondycji. Jedną z prostych, powszechnie stosowanych strategii jest algorytm zachłanny. To podejście oznacza rozważenie możliwych, natychmiastowych alternatyw i wybranie tej, która wydaje się przybliżać do celu. Na przykład w szachach postawienie przeciwnika w mata jest ostatecznym celem, ale zabranie hetmana

przeciwnika może zaprowadzić cię daleko. Strategia zachłanna to to, co badacze sztucznej inteligencji nazywają heurystyką, czyli regułą praktyczną lub skrótem, który często jest przydatny, ale nie ma gwarancji jego skuteczności. Ludzie również używają heurystyk. Badania wykazały, że rozwiązując problemy wyszukiwania, często utknęliśmy w punkcie, w którym musimy wykonać krok, który wydaje się oddalać nas od rozwiązania, mimo że może to być konieczny krok do rozwiązania całego problemu. Sztuczna inteligencja w grach zazwyczaj patrzy na wiele kroków naprzód, aby spróbować uniknąć tego problemu. Przeszukiwanie grafów za pomocą heurystyk działa dobrze, gdy każdy stan jest dobrze zdefiniowany i oddzielony od wszystkich pozostałych stanów, a liczba stanów jest skończona — nawet gdy, jak w szachach, liczba stanów może być zbyt duża, aby można ją było precyzyjnie obliczyć. W innych rodzajach problemów, takich jak lądowanie łazika marsjańskiego, możliwe stany, takie jak pozycja i orientacja robota, oraz działania, takie jak siła ciągu lub sterowania, są liczbami rzeczywistymi. W tym kontekście, gdzie istnieje rzeczywiście nieskończona liczba możliwych stanów, poszukiwanie jest często nazywane optymalizacją. W tym przypadku dobre rozwiązanie odpowiada punktowi lub regionowi wewnątrz wielowymiarowej przestrzeni utworzonej przez wszystkie zmienne stanu. Łatwo to sobie wyobrazić w przypadku jedno-, dwu- lub trójwymiarowej przestrzeni stanów , ale te same zasady obowiązują nawet w przypadku wielu wymiarów. W problemach optymalizacyjnych proces oceny jest często opisywany jako zastosowanie funkcji wartości, funkcji celu lub funkcji dopasowania. Wszystkie te terminy oznaczają mniej więcej to samo – opisują sposób oceny różnych stanów systemu w porównaniu z docelowym poziomem wydajności. W tym przypadku możemy również zastosować strategię zachłanną, która w problemach ciągłej optymalizacji nazywana jest wspinaczką górską. Opiera się ona na analogii, w której poszukiwanie jest jak poruszanie się po fizycznym krajobrazie, gdzie celem jest osiągnięcie najwyższego szczytu. Ocena obliczana dla bieżącej lokalizacji to aktualna wysokość w krajobrazie, a strategia wspinaczki górskiej polega po prostu na rozejrzeniu się w okolicy, sprawdzeniu, w którym kierunku jest góra, i skierowaniu się w tamtą stronę. Można sobie wyobrazić, że robisz coś podobnego, wybierając się na spacer po zboczu wzgórza w bardzo mglisty dzień. Ponownie, wspinaczka górska jest heurystyką, ponieważ podążanie za lokalnym nachyleniem może doprowadzić do lokalnego szczytu lub lokalnego optimum, a nie do najwyższego punktu w krajobrazie. To jeden z powodów, dla których wspinacze górscy gubią się w złą pogodę. Jest to również bardzo dobrze znany problem w sztucznej inteligencji, który może uniemożliwić algorytmom znalezienie najlepszych rozwiązań. Czasami problem sztucznej inteligencji jest przedstawiany jako spadek gradientu, a nie wspinaczka. To ta sama strategia, ale tym razem twierdzisz, że najlepszymi rozwiązaniami są najgłębsze doliny, a nie najwyższe szczyty, więc musisz schodzić w dół, aby znaleźć lepsze rozwiązania. Zawsze możesz przekształcić problem wspinaczki górskiej w problem spadku gradientu, zmieniając znak funkcji oceny (zamieniając znak dodatni na ujemny i odwrotnie), więc tak naprawdę są to te same rzeczy. W przypadku niektórych typów problemów bardziej intuicyjne jest mówienie o maksymalizacji czegoś, takiego jak dokładność, nagroda lub kondycja, a w przypadku innych problemów – o minimalizacji czegoś, takiego jak czas, energia, koszt lub błąd. Aby uniknąć lokalnych optimów w przestrzeni rozwiązań – czyli znalezienia szczytu, ale nie najwyższego – algorytmy optymalizacji wykorzystują różne strategie, takie jak posiadanie wielu pozycji startowych. Oznacza to, że można badać wiele możliwych ścieżek do celu, a nie tylko jedną. Algorytmy genetyczne inspirowane naturalnymi procesami leżącymi u podstaw ewolucji wykorzystują to podejście, zaczynając od populacji osobników, z których każdy jest reprezentowany przez inny sztuczny chromosom — często po prostu ciąg jedynek i zer. Każdy chromosom jest następnie używany do stworzenia potencjalnego rozwiązania problemu, na przykład poprzez wykorzystanie go do zbudowania sieci neuronowej. To rozwiązanie jest następnie testowane i oceniane na podstawie jego sukcesu w danym problemie. Te rozwiązania, które działają najlepiej, są następnie kopiowane do następnego pokolenia, często przy użyciu operatorów takich jak krzyżowanie (łączenie sekcji z dwóch lub więcej chromosomów) i mutacja (losowe odwracanie bitów), które ponownie są bezpośrednio inspirowane procesami występującymi

w ewolucji biologicznej. Z czasem członkowie każdej kolejnej populacji powinni stawać się lepsi w zadaniu, ostatecznie osiągając wyniki bliskie szczytom krajobrazu. W latach 90. informatyk Karl Sims wykorzystał algorytmy genetyczne do ewolucji populacji prostych, przegubowych stworzeń sterowanych przez sztuczne sieci neuronowe i poruszających się w symulowanych światach fizycznych. Sims wykazał, że może ewoluować stworzenia, aby były dobre w zadaniach związanych z poruszaniem się, takich jak pływanie w wodzie czy czołganie się po twardym gruncie. Odkrył również, że może odtworzyć ewolucyjny „wyścig zbrojeń", wystawiając jedną populację stworzeń przeciwko drugiej w zadaniu takim jak walka o kontrolę nad zasobem. Niektóre stworzenia Simsa wyglądały i poruszały się w sposób przypominający proste zwierzęta, ilustrując potęgę ewolucji i dając dowód na możliwość istnienia „sztucznego życia". Inną strategią ucieczki od lokalnych optimów jest po prostu dodanie szumu, czyli jittera, do reguły decyzyjnej, tak aby czasami, zupełnie przypadkiem, wykonać krok, który wydaje się oddalać od celu. Dodawanie szumu do procesu decyzyjnego wydaje się raczej nieinteligentne. Jednakże, jeśli próbujesz stać się w czymś lepszym, dodanie pewnego losowego odchylenia od tego, co mógłbyś zrobić w innych okolicznościach, może być dobrym sposobem na zbadanie przestrzeni pobliskich alternatywnych decyzji. Dowody z neuronauki sugerują, że biologiczne sieci neuronowe również mogą być zaszumionymi decydentami. Duże modele językowe, takie jak ChatGPT firmy Open AI, często dodają znaczący element losowości do wyboru kolejnych słów, ponieważ bez szumu ich rozwiązania mogą wydawać się powtarzalne i banalne. Przeszukiwanie i optymalizacja to tak potężne narzędzia, że kuszące może być ujmowanie w ten sposób wszystkich inteligentnych zachowań. Narzędzia te mają jednak swoje ograniczenia. Na przykład, rozumiejąc świat, często próbujemy identyfikować cechy, obiekty, ludzi i wzorce. W tych dziedzinach nie zawsze jest oczywiste, jaką miarę próbujemy zmaksymalizować (lub zminimalizować). W przypadku trudnych problemów najtrudniejszym zadaniem może być stworzenie odpowiedniego opisu przestrzeni problemu, w której można następnie zastosować przeszukiwanie lub optymalizację. Czasami w sztucznej inteligencji naukowcy badają problem i mogą opracować bezpośrednie rozwiązanie za pomocą sprytnych obliczeń matematycznych lub programowania. Na przykład, jeśli próbujesz poruszyć robota i wiesz wystarczająco dużo o jego siłownikach i stawach, możesz zapisać równania, znane jako równania kinematyczne, aby dokładnie określić, jak powinien się on poruszać, aby dotrzeć do określonego punktu w przestrzeni – bez konieczności wyszukiwania. Ogólnie rzecz biorąc, takie podejścia mogą prowadzić daleko i szybko, choć może być konieczne zastosowanie optymalizacji w celu dostrojenia niektórych parametrów. Zazwyczaj opisujemy takie podejścia jako inżynierię, a nie sztuczną inteligencję; Jednak tak naprawdę nie ma tu wyraźnego rozróżnienia. Jeśli ludzki umysł jest naszym modelem inteligencji, to mógł również nabyć niektóre z tych sprytnych, bezpośrednich rozwiązań.

ROZUMOWANIE PROBABILISTYCZNE

Rozważmy przykład codziennego rozumowania. Załóżmy, że wchodzisz do pokoju na przyjęciu akurat w momencie, gdy ktoś inny wychodzi innymi drzwiami. Dostrzegasz odchodzącą postać i zastanawiasz się, czy to może być twój dobry przyjaciel Joe. Postać była wysoka i miała długie, ciemne włosy, tak jak Joe. Wyglądała na ubraną w ciemną skórzaną kurtkę, co, jak wiesz, jest jednym z ulubionych elementów ubioru Joe. Rozglądając się po pokoju, widzisz Sally, o której wiesz, że jest również przyjaciółką Joe. Każda z tych informacji wskazuje na możliwość, że osoba, którą widziałeś, mogła być Joe, ale nic nie daje ci pewności. Możesz powiedzieć, że wskazówki sumują się i sprawiają, że jest więcej niż prawdopodobne, że to był Joe, ale jak dokładnie twój umysł to sumuje? To, co robi twój umysł, to niemal na pewno jakaś forma rozumowania probabilistycznego i jest w tym zadziwiająco dobry, nawet jeśli nie uważasz się za utalentowanego matematycznie. W niemal każdej sytuacji, w której się znajdujemy, mamy do czynienia z niepełnymi lub niepewnymi informacjami, a nasz umysł musi jak najlepiej zgadywać, co się dzieje. Możesz być świadomy części tego rozumowania, ale często dzieje się

ono w częściach umysłu, do których nie masz świadomego dostępu. Widzimy, naszym świadomym „okiem umysłu", najbardziej prawdopodobną interpretację świata, a nie wszystkie możliwości, które może rozważać podświadomość. Niektóre iluzje optyczne, takie jak iluzja kaczki i królika oraz sześcian Neckera, ilustrują tę niejednoznaczność. Analizując te obrazy, umysł przeskakuje między różnymi interpretacjami, ale nigdy nie jesteś w stanie świadomie dostrzec obu naraz. XIX-wieczny fizyk i psycholog Hermann von Helmholtz opisał to, co dzieje się w tych częściach mózgu, do których nie mamy świadomego dostępu, jako formę „nieświadomego wnioskowania". Teoria prawdopodobieństwa to gałąź matematyki, która opiera się na logice, ale łączy ją z teorią liczb. Jej początki sięgają XVI wieku, kiedy to matematyk Blaise Pascal i Pierre de Fermat wypowiedzieli się na temat gry w kości. Ich wysiłki położyły podwaliny pod badania nad niepewnością. Jednak prace XVII-wiecznego wikariusza, Thomasa Bayesa, wywarły prawdopodobnie największy wpływ na rozumowanie probabilistyczne w dziedzinie sztucznej inteligencji. Bayes, który za życia nie opublikował żadnych prac matematycznych, zajmował się prawdopodobieństwami warunkowymi – prawdopodobieństwem, że coś się wydarzy, na podstawie tego, że coś innego już się wydarzyło. Jego pomysł doprowadził do powstania tego, co obecnie nazywamy regułą Bayesa9 lub wnioskowaniem bayesowskim, metodą łączenia ze sobą obliczeń prawdopodobieństw w sposób podobny do tego, który rozważano wcześniej dla łączenia sekwencji dedukcji lub reguł jeśli-to. W tym przypadku jednak zamiast propagować proste twierdzenia o tym, co jest prawdą, a co fałszem, propagujemy teraz prawdopodobieństwa, czyli oczekiwania dotyczące prawdopodobieństwa, że coś jest prawdą (lub fałszem). Łączenie i propagowanie prawdopodobieństw wymaga wielu obliczeń, więc dopiero wraz z pojawieniem się wydajnych komputerów podejścia bayesowskie naprawdę się rozwinęły. W latach 80. XX wieku informatyk Judea Pearl10 wzbudził zainteresowanie rozumowaniem bayesowskim, łącząc je z teorią grafów, aby pokazać, jak wnioskowanie probabilistyczne może być propagowane poprzez sieć zmiennych w celu oszacowania prawdopodobieństwa stanów będących przedmiotem zainteresowania. W ten sposób, zaczynając od znanych prawdopodobieństw, które można zmierzyć lub oszacować, można dowiedzieć się, jak prawdopodobne są rzeczy, na których nam szczególnie zależy. Na przykład w systemach diagnostyki medycznej niektóre węzły mogą reprezentować wyniki testów i objawy, a inne diagnozy. Propagacja prawdopodobieństw w sieci może określić, które diagnozy są bardziej prawdopodobne niż inne – co ma realne znaczenie dla lekarzy i ich pacjentów. Obecnie wnioskowanie bayesowskie jest jedną z podstawowych technologii stosowanych w wielu systemach sztucznej inteligencji i kluczowym elementem teorii ludzkiej percepcji i inteligencji, które opierają się na koncepcji wnioskowania nieświadomego Helmholtza. Na przykład, niedawne badanie przeprowadzone przez neurobiologa Williama Arrisona i jego współpracowników wykazało, że aktywność mózgu związana z kodowaniem orientacji wzrokowej była lepiej dopasowana przez bayesowski model statystyczny niż alternatywne schematy kodowania. To dowód na to, że nieświadomość rzeczywiście może myśleć w kategoriach prawdopodobieństwa. Mówiąc szerzej, znaczna część dzisiejszej sztucznej inteligencji wykorzystuje metody oparte na statystycznych (tj. probabilistycznych) podejściach do znajdowania wzorców w danych. Dotyczy to wielu podejść wykorzystujących sztuczne sieci neuronowe, gdzie można wykazać, że zachowanie sieci jest zbliżone do znanej metody statystycznej. Na przykład, niektóre formy uczenia bez nadzoru działają, ponieważ przybliżają statystyczną metodę analizy czynnikowej, po raz pierwszy opisaną przez Charlesa Spearmana i omówioną w rozdziale 2 jako sposób identyfikacji różnych aspektów inteligencji na podstawie wyników testów IQ. Ta wiedza otwiera drzwi do zrozumienia biologicznych sieci neuronowych, które tworzą mózg, jako potężnych form przetwarzania statystycznego.

TEORIA STEROWANIA

Pamiętasz, że zanim pojawiła się sztuczna inteligencja, istniała cybernetyka? Zanim powstały potężne komputery cyfrowe, grupa badaczy zwanych cybernetykami wykorzystywała narzędzia teorii

sterowania, aby spróbować zrozumieć umysł i skonstruować inteligentne maszyny. Jedną z kluczowych idei teorii sterowania jest ujemne sprzężenie zwrotne. Rozważ termostat w swoim domu. To prosta forma regulatora z ujemnym sprzężeniem zwrotnym. Załóżmy, że ustawisz go na preferowaną temperaturę 21°C. W teorii sterowania nazywa się to nastawą. Jeśli temperatura w pomieszczeniu stanie się zbyt wysoka, powiedzmy 23°C, różnica między temperaturą rzeczywistą a preferowaną wyniesie +2°C. Termostat wykorzystuje ujemne sprzężenie zwrotne, więc używa wartości ujemnej tej liczby, -2°C, aby zdecydować, co zrobić, czyli czy ogrzać, czy zmrozić pomieszczenie. Ta wartość jest znana jako błąd. Twój termostat to bardzo prosty regulator, który ma tylko dwie możliwości — włączyć lub wyłączyć ogrzewanie. W tym przypadku chce obniżyć temperaturę w pomieszczeniu o -2°C, więc ogrzewanie zostaje wyłączone. Dwadzieścia minut później temperatura spadła i wynosi teraz 18°; różnica wynosi -2°, ale (ujemne) sprzężenie zwrotne jest tym razem dodatnie (dwa minusy dają plus), więc ogrzewanie jest ponownie włączone. Można się obawiać, że tak proste urządzenie będzie ciągle włączać i wyłączać ogrzewanie, co może zmniejszyć efektywność energetyczną lub zwiększyć zużycie. I masz rację. Aby uniknąć tego rodzaju oscylacji, znanych jako „polowanie", termostat ma zakres temperatur wokół wartości zadanej, powiedzmy plus lub minus 1°, znany jako zakres histerezy, w którym nie podejmie żadnych działań. Jeśli to się sprawdzi w przypadku ogrzewania domu, może to również pomóc wyjaśnić, jak mózg kontroluje niektóre ważne zmienne fizjologiczne niezbędne do życia, które również chcemy utrzymywać w pewnym zakresie. Należą do nich takie rzeczy, jak temperatura ciała, poziom glukozy we krwi, nawodnienie i ciśnienie krwi, żeby wymienić tylko kilka. Działanie polegające na utrzymywaniu krytycznych zmiennych życiowych w odpowiednim zakresie nazywa się homeostazą i było ważną częścią inspiracji dla cybernetycznego urządzenia Rossa Ashby'ego, Homeostatu. Wiele teorii homeostazy opiera się na koncepcji regulatora z ujemnym sprzężeniem zwrotnym. Teorie te zazwyczaj obejmują sterowanie ciągłe, a nie proste włączanie i wyłączanie, jak w przypadku termostatu. Sterowanie ciągłe może oznaczać, że siła sygnału sterującego jest proporcjonalna do wielkości błędu lub, aby zapewnić płynniejszą formę sterowania, sumy błędów w czasie, znanej jako sterowanie całkowe. Zatem, jeśli jest ci tylko trochę zimno, możesz poczuć, jak włosy stają ci dęba, co jest użytecznym mechanizmem zatrzymywania ciepła i odczuwania lekkiego dreszczu. Ale jeśli temperatura twojego ciała znacznie spadła, możesz poczuć, jak twoje mięśnie kurczą się znacznie intensywniej, próbując zwalczyć zimno. Sterowanie z ujemnym sprzężeniem zwrotnym można powiązać ze strategiami optymalizacji omówionymi wcześniej, ponieważ można je postrzegać jako formę gradientowego spadku w kierunku wartości docelowej. Pozostając przy problemie utrzymania ciepła, mózg, naturalnie, ma wiele strategii. Oprócz jeżenia włosów na ciele i drżenia, możesz zacząć spalać tkankę tłuszczową. Wszystkie te mechanizmy są regulowane w mózgu poniżej poziomu świadomości – nie możesz wybrać, żeby nie drżeć. Jednak gdy zauważysz, że drżysz, możesz zdecydować się na założenie grubszych ubrań lub przeniesienie się do cieplejszej części domu. Są to świadome decyzje, wymagające znacznie większego zaangażowania mózgu, w tym części kory mózgowej, do planowania i realizacji. Mimo to niektóre teorie psychologiczne nadal postrzegają te zachowania jako napędzane przez formy kontroli zwrotnej. Jedna z wpływowych teorii, opracowana w latach 50. XX wieku przez fizyka medycznego Williama Powersa, zakładała, że wszystkie zachowania są generowane przez hierarchię pętli sterujących, przy czym prostsze pętle napędzają zachowania takie jak drżenie, a bardziej złożone – poszukiwanie wełnianego swetra lub czapki. Ciało ludzkie jest skutecznym regulatorem temperatury i, gdy jest zdrowe, utrzymuje ją na stałym poziomie 37°C. Czy ten punkt nastawczy jest reprezentowany gdzieś w mózgu? Zaskakująco, odpowiedź brzmi: prawdopodobnie nie. Zamiast tego, różne mechanizmy utrzymywania temperatury mogą działać w pewnym stopniu niezależnie, wykorzystując różne mechanizmy kontrolne, sensoryczne i motoryczne, a być może każdy z nich ma swój własny zakres aktywnej temperatury. Lepiej byłoby wyobrazić sobie tę stałą temperaturę ciała na poziomie 37°C jako „punkt równowagi", który powstaje w wyniku interakcji wszystkich układów mózgu i organizmu odpowiedzialnych za regulację temperatury. Pojęcie

punktu nastawy wiąże się z innym problemem. Niektóre zmienne, które organizm musi kontrolować, nie mają stabilnej wartości docelowej; zmieniają się one w zależności od aktywności. Na przykład ciśnienie krwi zmienia się w ciągu dnia i musi być wyższe podczas intensywnej aktywności, takiej jak ćwiczenia fizyczne, niż w spoczynku. Zmienne fizjologiczne, takie jak ciśnienie krwi, zmieniają się w sposób, który przewiduje zapotrzebowanie, co wskazuje na formę kontroli predykcyjnej. Nawet temperatura ciała może się zmieniać. Na przykład, gdy walczysz z infekcją, temperatura ciała wzrośnie, ponieważ układ odpornościowy może działać skuteczniej, gdy ciało jest nieco cieplejsze. Zwróciliśmy uwagę na różne historyczne korzenie teorii systemów inteligentnych wywodzących się z informatyki i teorii sterowania. Wcześni teoretycy sztucznej inteligencji chcieli odróżnić swoją dziedzinę od cybernetyki, dlatego przyjęli nową etykietę. Nawet dzisiaj badacze sztucznej inteligencji i teoretycy sterowania prawdopodobnie pracują na różnych wydziałach uniwersyteckich. Jednak błędem byłoby pomijanie narzędzi dostarczanych przez teorię sterowania do zrozumienia ludzkiego umysłu lub tworzenia sztucznej inteligencji. Rzeczywiście, gdy pracujemy ze złożonymi artefaktami sztucznej inteligencji, takimi jak roboty, narzędzia matematyczne teorii sterowania odgrywają kluczową rolę w sposobie, w jaki sprawiamy, że maszyny te utrzymują się w pozycji pionowej, poruszają się i, bardziej ogólnie, wykazują inteligentne zachowania.

CZY LUDZIE MYŚLĄ INACZEJ?

Istnieją obszary ludzkiego rozumowania i myślenia, które uważa się za wykraczające poza sztuczną inteligencję. Zakończę ten rozdział omówieniem kilku bardziej znanych przykładów. Zacznijmy od rozumowania analogicznego. Jest to umiejętność rozwiązywania problemów poprzez porównywanie ich z sytuacjami napotkanymi wcześniej, być może w zupełnie innej dziedzinie. Przykładami odkryć naukowych, o których mówi się, że były wynikiem rozumowania analogicznego, są odkrycie struktury cząsteczki DNA przez Francisa Cricka i Jamesa Watsona oraz szczególna teoria względności Alberta Einsteina. W obu przypadkach postęp nastąpił dzięki „myśleniu nieszablonowemu" i analizie innych problemów z chemii i fizyki, a w przypadku Cricka i Watsona – budowaniu modeli fizycznych w celu osiągnięcia postępu. Myślenie analogiczne wydaje się polegać na dostrzeganiu wzorców i podobieństw między sytuacjami, które w innym przypadku mogłyby wydawać się niepowiązane. W kolejnym rozdziale przekonamy się, że abstrakcje są osiągalne w sztucznej inteligencji i jest to kluczowe dla sposobu, w jaki rozwiązują one problemy. Dopasowanie abstrakcji w różnych dziedzinach wciąż stanowi wyzwanie dla sztucznej inteligencji, ale poczyniono postęp w tym kierunku, na przykład w dziedzinie sztucznej inteligencji generującej tekst na obraz. W przypadku fizycznego modelu DNA Cricka i Watsona, wykonanego z kawałków drutu i metalu, możemy uznać to za przykład „myślenia poza głową", czyli tego, co filozof Andy Clark nazwał „rozszerzonym umysłem". Jest to zdolność ludzi do wykorzystywania otoczenia do rozszerzania swoich zdolności myślenia – do codziennych przykładów należą długopis i papier, kalendarze, mapy, kalkulatory, telefony komórkowe i oczywiście komputery. Niewątpliwie Crick i Watson mieli również dużą wiedzę na temat najnowocześniejszych osiągnięć nauki w dziedzinie biologii, chemii i fizyki, co umożliwiło im dokonanie odkrycia, więc połączenie tej wiedzy z wykorzystaniem niektórych z tych zewnętrznych narzędzi doprowadziło do ich postępu. Albert Einstein był wysoce wizualnym myślicielem, który jako nastolatek wyobrażał sobie, że biegnie obok wiązki światła i wyobraża sobie, że fala jest stacjonarna. Einstein wykorzystywał również wyimaginowane scenariusze z udziałem pociągów dużych prędkości i spadających wind, aby uzasadnić swoje idee dotyczące teorii względności. Innym pokrewnym obszarem myślenia jest to, co filozofowie nazywają rozumowaniem abdukcyjnym. Abdukcja różni się od dedukcji i indukcji, ponieważ wydaje się polegać na rozumowaniu do najlepszego wyjaśnienia i „poza dowodami". W ludzkim rozumowaniu abdukcja może obejmować heurystykę, rozumowanie oparte na zdrowym rozsądku i to, co nazywamy intuicją. Jedną z powszechnie stosowanych heurystyk jest oszczędność, poszukiwanie najprostszego i najbardziej eleganckiego rozwiązania, znane również jako brzytwa Ockhama. Abdukcja przypomina

nieco rozumowanie probabilistyczne, ponieważ obie polegają na zgadywaniu najlepszego rozwiązania przy ograniczonych informacjach. Połączenie istniejących technologii sztucznej inteligencji i odpowiedniego wykorzystania heurystyk mogłoby umożliwić im rozumowanie abdukcyjne, choć trudno będzie tu wyciągnąć jednoznaczne wnioski, ponieważ psychologia i filozofia nie zdołały przedstawić jasnego wyjaśnienia, na czym polega abdukcja w ludzkim rozumowaniu. Szerzej rzecz biorąc, wiele mówi się o tym, że ludzkie podejmowanie decyzji jest motywowane czynnikami innymi niż czysta racjonalność. Z tą ideą zetknęliśmy się w rozdziale 2, gdzie poznaliśmy różne koncepcje inteligencji, w tym inteligencję emocjonalną i rozróżnienie Daniela Kahnemana między myśleniem „szybkim" a „wolnym". Szybkie myślenie zostało opisane jako automatyczne, emocjonalne, a często instynktowne i nieświadome, i przeciwstawione bardziej systematycznej, świadomej i logicznej naturze myślenia powolnego. To, że proces jest automatyczny i nieświadomy, nie wyklucza jego racjonalności. Jak już wcześniej omawiałem, nasz mózg rozwiązuje wiele problemów poniżej poziomu świadomości, ale można to postrzegać jako formę wnioskowania. Ale co z emocjami i instynktem? Jednym ze sposobów myślenia o tych formach wiedzy jest to, że są one skumulowaną mądrością nabytą poprzez ewolucję i uczenie się. Jeśli w przeszłości nabyłeś dobre rozwiązanie problemu, to dlaczego nie miałbyś go zapisać i wykorzystać ponownie? W najszerszym sensie jest to to, co ewolucja jest w stanie zrobić, buforując lub gromadząc użyteczne rozwiązania w formie projektów dla naszych ciał i mózgów. Jeśli wdrożysz ewolucyjne rozwiązanie problemu – to, co psychologowie nazwaliby instynktem – możesz być w stanie zrobić to, co słuszne, ale nie będziesz w stanie wyjaśnić, dlaczego. To, że możesz mieć „przeczucie" co do tego, co należy zrobić, może mieć związek z naturą emocji. Przynajmniej w jednym sensie są to oceny tego rodzaju, o których mówiliśmy już, omawiając wyszukiwanie i optymalizację. Instynktowne lub nabyte oceny mogą pozwolić ci bardzo szybko ocenić sytuację i zdecydować, czy jest ona dobra, czy zła. Neuropsycholog Antonio Damasio badał uszkodzenia mózgu w obszarach emocjonalnych, które prowadziły do podejmowania złych decyzji. Zaproponował „hipotezę markera somatycznego", zgodnie z którą obszary mózgu zaangażowane w odczuwanie ciała mogą przechowywać skojarzenia między sytuacjami a rezultatami, które są następnie doświadczane jako emocje. Istnieją dobrze poznane metody uczenia się tego rodzaju ocen, które zostały zbadane w dziedzinie sztucznej inteligencji w uczeniu się przez wzmacnianie i są obecnie wykorzystywane do trenowania niektórych modeli uczenia się na dużą skalę. Emocje wciąż owiane są pewną tajemnicą, ponieważ ludzkie doświadczanie emocji nie do końca pasuje do tej koncepcji emocji jako oceny – w końcu to może być tylko aktywacja kilku komórek mózgowych. Jednak w organizmach zwierząt (i ludzi) musimy zwrócić uwagę na ścisłe powiązanie między stanami mózgu i ciała. Ocena wygenerowana w mózgu doprowadzi do dalszych efektów, w tym do uwolnienia neuromodulatorów i odpowiedniego zachowania fizycznego – na przykład ucieczki w przypadku strachu. Doświadczenie emocji może zatem mieć wiele wspólnego z tą mieszanką oceny i reakcji na sytuację zewnętrzną, a następnie „odczuwania" reakcji własnego ciała. Intensywność tej reakcji prawdopodobnie zaangażuje o wiele więcej obwodów mózgowych, wzmacniając doznanie i czyniąc je niezapomnianym. Oprócz rodzajów skompilowanej inteligencji zapewnianych przez ewolucję i uczenie się, ludzkie rozumowanie kieruje się również heurystykami lub regułami praktycznymi. Jednak, jak widzieliśmy, heurystyki są również szeroko wykorzystywane przez sztuczną inteligencję. W rzeczywistości, nawet gdy dysponujemy kompletną teorią matematyczną dotyczącą tego, jak rozumowo rozwiązać problem i znaleźć idealne rozwiązanie, okazuje się, że wymagane obliczenia mogą być zbyt kosztowne, aby były opłacalne lub wykonalne w dostępnym czasie. W dużej mierze w sztucznej inteligencji (AI) chodzi zatem o znalezienie użytecznych przybliżeń, które prowadzą do użytecznego rozwiązania, przy jednoczesnym pomijaniu kilku technicznych zawiłości. Jeden z twórców AI, Herb Simon, którego poznaliśmy , zauważył, że dotyczy to również ludzi. Zasugerował, że mamy „ograniczoną racjonalność", co oznacza, że dysponujemy ograniczonymi zasobami poznawczymi, czas jest zazwyczaj ograniczeniem i zazwyczaj zadowala nas rozwiązanie, które jest wystarczająco dobre, a nie prawdziwie optymalne. Wreszcie, często zauważa

się, że ludzkie rozumowanie może być stronnicze, co oznacza, że nie przyjmujemy najbardziej racjonalnego rozwiązania problemu, czasami w sposób, który wydaje się sprzeczny z naszymi najlepszymi interesami. Na przykład Daniel Kahneman i jego współpracownik Amos Tversky opisali coś, co nazwali „awersją do strat". Jest to powszechna ludzka tendencja do preferowania unikania kary za poniesienie straty nad korzyścią w postaci równoważnego zysku. Awersja do strat może prowadzić do złych decyzji ekonomicznych. Na przykład niedoświadczeni inwestorzy często trzymają się akcji tracących na wartości, gdy lepszym rozwiązaniem może być ich sprzedaż, lub unikają inwestycji postrzeganych jako nadmiernie ryzykowne. Jednakże, chociaż awersja do strat może sprawić, że przeciętny człowiek stanie się kiepskim inwestorem, istnieją inne konteksty, w których można to uznać za racjonalną strategię. Na przykład, z perspektywy ewolucyjnej, straty mogą zagrażać życiu, a jeśli fundamentalnym celem jest przetrwanie, awersja do ryzyka może być lepszą strategią. Czy zatem naprawdę istnieje forma rozumowania lub myślenia, którą posiadają ludzie, a której SI mogą nigdy nie posiadać? Nie jestem pewien. Kiedy połączymy ze sobą SI na dużą skalę, być może z kilkoma systemami składowymi opisanymi w tym miejscu, dojdziemy do form myślenia, które bardziej przypominają rozumowanie analogowe lub abdukcję. Myślę, że zobaczymy również SI, które będą miały odpowiednik szybkiego i wolnego myślenia – będą zdolne do podejmowania szybkich i tanich decyzji w oparciu o rozpoznawanie wzorców i oceny zapisane w procesie uczenia się, ale będą mogły następnie posłużyć się wolniejszymi procesami dedukcji i rozumowania, aby zweryfikować, czy decyzja była słuszna. Zbadamy to szerzej w kolejnych rozdziałach.

UCZENIE SIĘ W SIECIACH NEURONOWYCH

Sztuczna inteligencja rozwija się falami, z spokojniejszymi okresami pomiędzy nimi, które poetycko określa się mianem „zim AI". Pod koniec XXI wieku i na początku XXI wieku nowa fala zaczęła nabierać tempa. Był to początek ery modeli wielkoskalowych sztucznych sieci neuronowych, popularnie zwanych głębokimi sieciami neuronowymi lub po prostu głębokim uczeniem. Tego rodzaju SI przyciągnęły uwagę opinii publicznej i nie bez powodu. W wielu dziedzinach, w których ludzie do tej pory przewyższali SI, maszyny doganiały, a nawet przewyższały swoich ludzkich odpowiedników. Dotyczy to szczególnie dziedzin, które klasycznie nie były postrzegane jako wymagające dużej inteligencji – naszej zdolności do rozumienia świata za pomocą zmysłów, znanej również jako percepcja. Przypomnijmy sobie, jak Arystoteles rozróżniał percepcję (aesthesis) od intelektu (nuos). Ta idea prymatu intelektu do dziś utrzymuje się w niektórych obszarach sztucznej inteligencji, ale jak przyjrzymy się w tym rozdziale, w przypadku sieci neuronowych droga do zrozumienia i odtworzenia inteligencji na poziomie ludzkim prowadzi przez percepcję, a szerszym odkryciem jest to, że to, co sprawdza się w przypadku percepcji, może działać również w przypadku intelektu.

OD PERCEPCJI DO PERCEPTRONU

W początkach informatyki dostępne technologie nadawały się bardziej do badania procesów rozumowania niż do radzenia sobie ze światem rzeczywistym. Pierwsze komputery cyfrowe zaczęły pojawiać się na uniwersytetach pod koniec lat 40. XX wieku, ale ówczesne urządzenia sensoryczne – mikrofony i pierwsze kamery telewizyjne – były nadal analogowe. Ich wyjściami były sygnały elektryczne, które musiały zostać przekształcone na sygnały cyfrowe, zanim mogły zostać przetworzone przez komputery. Kiedy w latach 60. XX wieku rozpoczęły się badania nad widzeniem komputerowym, naukowcy digitalizowali czarno-białe rysunki liniowe, ponieważ praca z kamerami była po prostu zbyt trudna. W latach 50. XX wieku nauka o tym, jak mózg przetwarza sygnały sensoryczne, również dopiero się rozwijała. Neurobiolodzy David Hubel i Torsten Wiesel zarejestrowali w pierwotnej korze wzrokowej znieczulonego kota, wyświetlając mu przed oczami proste wzory światła. Znaleźli pojedyncze neurony korowe, które były szczególnie wrażliwe na paski światła prezentowane w różnych orientacjach, co sugeruje obecność komórek detekcji krawędzi w mózgu. Był

to przydatny dowód na to, że mózg nie przechwytuje bezpośredniej kopii tego, co widzimy oczami, a następnie nie prezentuje tego umysłowi na pewnego rodzaju wewnętrznym ekranie. Zamiast tego mózg zaczyna dekonstruować bodziec sensoryczny od samego początku, wybierając przydatne elementy, które mogą informować o tym, co dzieje się na świecie i co robić dalej. To, że neurony mogą działać jako detektory cech, aby znajdować krawędzie w scenie wizualnej, a następnie znajdować i rozpoznawać obiekty i tak dalej, jest częścią tego procesu. Mniej więcej w tym samym czasie, gdy Hubel i Wiesel dokonywali swoich odkryć, młody inżynier o nazwisku Frank Rosenblatt zbudował pierwsze fizyczne urządzenie do utworzenia sztucznej sieci neuronowej. Znana jako perceptron (połączenie słów percept-ion i neu-ron), maszyna Rosenblatta połączyła warstwę elektronicznych, binarnych urządzeń wejściowych z „neuronem" decyzyjnym, w którym każde wejście było połączone z decydentem za pomocą modyfikowalnego, ważonego połączenia o wartościach rzeczywistych. Decydent miał również modyfikowalny próg; jego wyjście było równe jeden, jeśli ważona suma wejść była większa od tego progu i zero w przeciwnym wypadku. Innymi słowy, podejmował binarną decyzję tak/nie. Rosenblatt wykorzystał regułę uczenia się z korekcją błędów, znaną obecnie jako reguła perceptronu, do trenowania swojego urządzenia w zakresie problemów rozpoznawania wzorców, takich jak rozpoznawanie odręcznych liter i cyfr. Chociaż późniejsze urządzenia zawierały system obrazowania, we wczesnych prototypach wzorce były wprowadzane bezpośrednio do komputera za pomocą perforowanych kart. Na przykład, jeśli celem było rozpoznanie litery X, perceptronowi przedstawiono by wiele odręcznych wersji litery X wraz z przykładami innych liter. Reguła uczenia Rosenblatta wykorzystywała różnicę między docelowym wynikiem a rzeczywistym wynikiem. Istnieją dwa sposoby, w jakie perceptron mógłby błędnie sklasyfikować obraz. Po pierwsze, jeśli obraz nie był literą X i odpowiedziałby jedynką (tak), wówczas sygnał błędu byłby ujemny (0−1 = −1). Po drugie, jeśli obraz był literą X i odpowiedziałby zerem (nie), wówczas sygnał błędu byłby dodatni (1−0 = +1). Jeśli klasyfikacja byłaby poprawna, błąd wynosiłby zero, a wagi nie uległyby zmianie. Stosując tę regułę uczenia, perceptron stopniowo poprawiałby swoje zadanie, wzmacniając wagi detektorów cech, które konsekwentnie aktywowały się, gdy X był obecny, i osłabiając wagi detektorów, które aktywowały się, gdy X był nieobecny. Z czasem błędów i sygnałów uczenia się niezerowych byłoby coraz mniej. Gdy perceptron poprawnie klasyfikował wszystkie obrazy, wagi w ogóle by się nie zmieniały. W terminach wyszukiwania można by powiedzieć, że zakończył on proces spadku gradientu. W 1959 roku dwaj inżynierowie elektrycy, Bernard Widrow i Ted Hoff, odkryli nową regułę uczenia się, znaną obecnie jako reguła delty, którą można traktować jako uogólnienie reguły uczenia się perceptronu. Nowa reguła nie ograniczała się do binarnych wejść i wyjść, a minimalizowała sumę kwadratów błędów dla wszystkich wzorców w zbiorze danych w sposób, który można było powiązać z metodami stosowanymi przez statystyków do wyszukiwania wzorców w danych. Oznaczało to, że oprócz wiedzy o działaniu reguły uczenia się istniało dobre wyjaśnienie, dlaczego ona działa. Widrow i Hoff opracowali również sprzętowe urządzenie klasyfikacyjne o nazwie Adaline, które implementowało ich model uczenia się. Obecnie modele takie jak Perceptron i Adaline są opisywane jako należące do klasy jednowarstwowych sieci neuronowych i postrzegane jako pionierskie kroki w kierunku potężnych sieci neuronowych AI XXI wieku. Pod koniec lat 50. XX wieku AI budziła duże zainteresowanie. Na przykład Perceptron Rosenblatta został okrzyknięty przez „New York Times" „elektronicznym mózgiem, który „uczy się" sam". Drugi artykuł w „Timesie", oparty na konferencji prasowej z udziałem samego Rosenblatta, mówił o komputerze elektronicznym, który „byłby w stanie chodzić, mówić, widzieć, pisać, reprodukować się i być świadomym swojego istnienia". Jednak w ciągu dekady badacze sztucznej inteligencji odwrócili się od obietnicy sieci neuronowych i powrócili do mniej inspirowanych mózgiem form sztucznej inteligencji. Powód? Jednowarstwowe sieci neuronowe nie są w stanie rozwiązać wielu rodzajów problemów klasyfikacyjnych, w tym tych stosunkowo łatwych dla ludzi i obejmujących podstawowy układ logiczny znany jako wyłączny-lub. To ograniczenie zostało szeroko omówione w słynnej książce „Perceptrons", opublikowanej w 1969 roku przez Marvina Minsky'ego i jego kolegę

Seymoura Paperta. Minsky i Rosenblatt chodzili razem do liceum, ale podobno byli rywalami, co może częściowo tłumaczyć entuzjazm Minsky'ego dla rozwiania bańki Rosenblatta. W 1972 roku kolejna wpływowa książka filozofa Huberta Dreyfusa zatytułowana „Czego komputery nie potrafią", przypuściła ostry atak na sztuczną inteligencję, krytykując jej zdolność do pozyskiwania ludzkiej wiedzy eksperckiej. Dreyfus argumentował, że ludzka wiedza jest kontekstowa, często ukryta i intuicyjna, a zatem trudna do wyjaśnienia lub uchwycenia za pomocą programów komputerowych. Odrzucał możliwość, że komputery cyfrowe mogłyby być przydatne do zrozumienia mózgu, i nie doceniał podejścia Rosenblatta do modelowania sieci neuronowych przypominających mózgi. Dreyfus zauważył w szczególności, że ludzie są ekspertami w dostrzeganiu wzorców w danych, w tym w naszym codziennym doświadczeniu percepcyjnym, a rozpoznawanie wzorców wykracza poza fizyczne podobieństwo i jest silnie zależne od kontekstu. Dla Dreyfusa sztuczna inteligencja była czysto bierna i odcieleśniona, podczas gdy ludzka inteligencja była aktywna, ucieleśniona i zaangażowana. Percepcja była holistyczną czynnością, w której „znaczenie szczegółów" było określane równolegle do znaczenia całości.

UCZENIE MASZYNOWE I UCZENIE SIĘ ZWIERZĄT

Przed rozwojem sztucznej inteligencji (AI), a także równolegle z nim, psychologowie próbowali zrozumieć, dlaczego zwierzęta tak skutecznie się uczą. Na przykład eksperymenty behawiorysty B. F. Skinnera wykazały, że gołębie można wyszkolić do dziobania celów w celu otrzymania nagrody w postaci jedzenia, a także do dokonywania złożonych osądów na temat bodźców wzrokowych, w tym do rozróżniania kształtów geometrycznych i abstrakcyjnych wzorów. Skiner był pod tak wielkim wrażeniem tych zdolności, że gołębie rozważano jako potencjalny środek sterowania bombami kierowanymi podczas II wojny światowej. Pomysł polegał na tym, że gołąb, wyszkolony w rozpoznawaniu konkretnego celu wojskowego na podstawie zdjęć, siedziałby w stożku dziobowym bomby i dziobał czuły ekran, gdy cel się pojawił, kierując bombę w stronę celu. Bomby kierowane przez gołębie okazały się niepraktyczne i nigdy nie zostały użyte w akcji, ale fakt, że w ogóle je rozważano, dowodzi mocy uczenia się wzrokowego u zwierząt. W 1972 roku psychologowie Robert Rescorla i Allan Wagner zaproponowali regułę uczenia się, która odegrała ogromną rolę w wyjaśnieniu wielu zjawisk związanych z uczeniem się zwierząt, a następnie została rozszerzona na ludzką percepcję, uczenie się kategorii i rozumowanie. Reguła Rescorli-Wagnera, jak jest obecnie znana, została wkrótce uznana za formalnie równoważną regule delty Widrowa-Hoffa, co ilustruje powszechność tego potężnego mechanizmu uczenia się zarówno u zwierząt, jak i u maszyn. Według Rescorli i Wagnera, uczenie się zachodzi zawsze, gdy istnieje rozbieżność między przewidywanymi a rzeczywistymi wynikami (jak miało to miejsce również w przypadku reguły uczenia się Rosenblatta). Jeśli zdarzenie jest zaskakujące w tym sensie, że nie było w pełni oczekiwane, następuje uczenie się. Jeśli zdarzenie jest w pełni przewidziane, uczenie się nie jest wymagane.

NOWY KONEKCJONIZM

Sceptycyzm generowany przez Dreyfusa i innych, w połączeniu z ograniczeniami istniejących systemów sztucznej inteligencji, doprowadził w latach 70. XX wieku do zjawiska, które obecnie nazywa się pierwszą zimą sztucznej inteligencji. Prawdę mówiąc, sztuczna inteligencja również padła ofiarą własnego szumu medialnego, ponieważ przewidywania Minsky'ego, Simonsa, Rosenblatta i innych nie spełniły się. Sytuacja ta jednak nie trwała długo. W latach 80. XX wieku zainteresowanie sztuczną inteligencją opartą na wiedzy odżyło wraz z pojawieniem się pierwszych komercyjnych systemów eksperckich. W tym okresie szeroko eksponowano również dzieła sztuki stworzone przez AARON, sztuczną inteligencję opracowaną przez byłego programistę Harolda Cohena, rozbudzając ciekawość dotyczącą możliwości sztucznej kreatywności. Zainteresowanie sztucznymi sieciami neuronowymi również zaczęło się odradzać, częściowo dzięki rosnącej interakcji między psychologami a badaczami

sztucznej inteligencji w ramach nowej interdyscyplinarnej dziedziny kognitywistyki. Jednym z najbardziej ekscytujących odkryć było odkrycie sposobu na pokonanie ograniczeń jednowarstwowych sieci neuronowych poprzez trenowanie sieci wielowarstwowej. Rosenblatt i pierwsi koneksjoniści wiedzieli, że sieć wielowarstwową można zaprogramować do rozwiązywania problemów nierozwiązywalnych w sieci jednowarstwowej. Problemem zawsze było to, jak zrobić to automatycznie poprzez uczenie się, czyli jak dostosować połączenia sieci po przejściu poza pierwszą warstwę adaptowalnych wag. W połowie lat 80. David Rumelhart i Robert Williams z nowo utworzonego Wydziału Kognitywistyki w San Diego w Kalifornii nawiązali współpracę z psychologiem Geoffreyem Hintonem, aby wykazać, że możliwe jest trenowanie wielowarstwowej sieci neuronowej z regułą uczenia się korygującą błędy, będącą uogólnieniem reguły delta. Nowa metoda uczenia się została nazwana propagacją wsteczną. Dokładniej, zaczynając od strony wyjściowej sieci, Rumelhart i współpracownicy wykazali, że można dostosować wyjściową warstwę wag tak, jak robi się to w sieci jednowarstwowej, a następnie, pracując wstecz w całej sieci, modyfikować sygnał błędu zgodnie z wartościami wag połączeń prowadzących z powrotem do każdej jednostki w warstwie poniżej. Aby to zadziałało, każdy sztuczny neuron (lub przynajmniej te w warstwach pośrednich) musiał obliczyć pewną nieliniową funkcję swojego wejścia. Oznacza to, że wyjście jednostki nie mogło być po prostu ważoną sumą swoich wejść (funkcją liniową); zamiast tego musiało zostać w jakiś sposób zniekształcone. Oznaczało to również, że warstwy pośrednie sieci – te, które nie były ani wejściowe, ani wyjściowe – nauczyły się interesujących sposobów reprezentowania wejść istotnych dla rozwiązania danego problemu. Innymi słowy, był to sposób na automatyczne projektowanie detektorów cech bez konieczności ręcznego kodowania. Propagacja wsteczna błędu szybko stała się znana jako propagacja wsteczna. Korzystając z metody backprop, badacze wykazali, że proste problemy, które Minsky i Papert uznali za niemożliwe do rozwiązania dla sieci jednowarstwowej, teraz mogą zostać rozwiązane. Co być może równie ważne, nowi „koneksjoniści" wykorzystali rosnącą moc i szybkość komputerów, aby wykazać, że sieci wielowarstwowe mogą uczyć się interesujących zadań, w których do tej pory tylko ludzie byli dobrzy. Na przykład czytanie na głos stanowiło wyzwanie dla programów AI, ponieważ wymowa, w języku takim jak angielski, nie jest fonetyczna. Nie ma prostych reguł, które odwzorowują pisownię na dźwięk, a nawet jeśli uda się znaleźć regułę, często istnieją wyjątki. Na przykład w języku angielskim „gave", „rave" i „save" mają podobną wymowę, ale „have" jest inne (i to bez oczywistego powodu). Oznaczało to, że czytanie na głos było trudne do uchwycenia za pomocą podejścia opartego na regułach. Model NETtalk Terry'ego Sejnowskiego i Charlesa Rosenberga, opublikowany w 1987 roku , nauczył się czytać i wymawiać niemal do poziomu biegłości dorosłego użytkownika. Po pełnym wytrenowaniu NETtalk osiągnął 95% dokładności wymowy w zestawie testowym kilkuset słów. Nieźle jak na model z około trzystoma sztucznymi neuronami, który nauczył się czytać od podstaw w zaledwie kilka dni. Co ciekawe, w miarę uczenia się NETtalk popełniał błędy podobne do tych popełnianych przez początkujących czytelników, co sugeruje, że sieci wielowarstwowe wykonywały czynności podobne do tych wykonywanych przez mózg.

TRZECIA FALA

Jak zauważono , kolejna fala entuzjazmu dla sztucznej inteligencji (AI) rozpoczęła się pod koniec XXI wieku, ponownie wywołana postępem w dziedzinie sieci neuronowych. W ciągu dwóch dekad od odkrycia propagacji wstecznej (backprop) nastąpił oczywiście rozwój coraz potężniejszych komputerów. Wraz z pojawieniem się internetu pojawiła się również ogromna ilość łatwo dostępnych, generowanych przez ludzi treści, które można było wykorzystać do trenowania sieci neuronowych. Akceleratory sprzętowe, takie jak procesory graficzne i farmy serwerów, które firmy technologiczne, takie jak Google i Facebook, budowały, aby wspierać rozwój internetu, umożliwiły badaczom budowanie i trenowanie coraz większych modeli z coraz większą ilością danych. Percepcja ponownie stała się kluczowym celem, ale teraz problem tłumaczenia z analogowego na cyfrowy zniknął. Cała

treść, zwłaszcza wideo i audio, znajdowała się w domenie cyfrowej, a część z nich w łatwych do manipulowania bazach danych. To był również początek ery „big data", a firmy zdawały sobie sprawę z przydatności posiadania danych, które były dobrze ustrukturyzowane i oznaczone, zazwyczaj przez ludzi, aby określić, czym jest dana treść. W obszarach takich jak rozpoznawanie mowy i widzenie komputerowe, wyzwania były na nowo definiowane w oparciu o te duże zbiory danych. Wyniki kodowania ludzkiego stanowiły złoty standard, a zadaniem było sprawdzenie, jak blisko można się zbliżyć. Inną ideą zapożyczoną z biologii była koncepcja splotowej sieci neuronowej (CNN). Jak wspomniano , widzenie jest przetwarzane w wielu obszarach mózgu, ale wspólną cechą jest to, że organizacja przestrzenna świata wizualnego jest w dużej mierze zachowana. Oznacza to, że pobliskie fragmenty przestrzeni są przetwarzane przez sąsiednie obszary siatkówki i tak samo przez całą drogę aż do warstw przetwarzania wzrokowego kory mózgowej. Ten typ organizacji, określany mianem topograficznej, sugeruje, że na każdym poziomie sąsiednie grupy komórek mogą wykonywać podobne obliczenia na danych wejściowych pochodzących z lokalnych obszarów w warstwie niższej. W sztucznej sieci splotowej pomysł ten jest realizowany za pomocą pojedynczego zestawu regulowanych wag, który jest stosowany do wszystkich lokalnych regionów sieci na każdej warstwie. Sieć porusza się po obrazie, region po regionie, w procesie znanym jako splot. Wykonanie tych samych obliczeń we wszystkich lokalnych obszarach obrazu pozwala zaoszczędzić na liczbie potrzebnych wag adaptacyjnych. Oznacza to również szybszą naukę, ponieważ wszystkie obszary obrazu mogą przyczyniać się do trenowania jednego zestawu wag. Od lat 80. XX wieku wprowadzono znaczące ulepszenia i modyfikacje algorytmów uczenia się, takich jak propagacja wsteczna, zwiększając ich wydajność. Obecnie lepiej rozumiemy, jak działają te algorytmy. Szersze pole uczenia maszynowego obejmuje również metody probabilistyczne, takie jak wnioskowanie bayesowskie. Trwa jednak debata na temat tego, czy sztuczna inteligencja dysponuje odpowiednimi algorytmami uczenia się. Chociaż propagacja wsteczna jest skuteczna, pojawia się pytanie, czy inny algorytm uczenia się mógłby być dokładniejszy, lepszy w generalizacji lub szybszy w uczeniu się? Okazuje się, że jednym z kluczowych przełomów było po prostu zbudowanie większych i głębszych sieci neuronowych, to znaczy takich, które zawierają zarówno więcej sztucznych neuronów, jak i więcej warstw; dwa osiągnięcia, które sprawiają, że sztuczne sieci neuronowe bardziej przypominają mózg. Kumulatywne efekty tych ulepszeń i możliwość trenowania większych i głębszych modeli stopniowo pojawiły się w latach 2000., ale nowe metody głębokiego uczenia się naprawdę trafiły na pierwsze strony gazet pod koniec tej dekady. Jednym z kroków naprzód około 2009 roku było to, że modele mowy wykorzystujące sieci neuronowe wyprzedziły te pracujące na innych zasadach. Kolejnym obszarem było widzenie komputerowe. W 2012 roku Alex-Net, głęboka sieć neuronowa uczenia się opracowana przez Alexa Krizhevsky'ego, Ilyę Sutskevera i Geoffreya Hintona (jednego z wcześniejszych odkrywców backprop), wygrała ważny coroczny konkurs rozpoznawania wzorców w dziedzinie widzenia komputerowego. Konkurs ten wymagał algorytmów do prawidłowej klasyfikacji tysięcy różnych obiektów, takich jak zwierzęta, budynki, żywność, artefakty stworzone przez człowieka oraz elementy naturalne, takie jak drzewa i chmury, i tak dalej. Zestaw treningowy był wstępnie oznakowany i składał się z milionów obrazów. Sieć AlexNet, to ośmiowarstwowa splotowa sieć neuronowa z 650 000 sztucznych neuronów i sześćdziesięcioma milionami parametrów. Podczas testów w konkursie w 2012 roku popełniła mniej niż połowę błędów w porównaniu z najbliższym rywalem.

CZY GŁĘBOKIE SIECI NEURONOWE NAPRAWDĘ PRZYPOMINAJĄ MÓZG?

Splotowe sieci neuronowe, takie jak AlexNet, są inspirowane mózgiem, ale czy przetwarzanie w mózgu jest w jakikolwiek sposób podobne do tego w głębokiej sieci neuronowej? Istnieją przesłanki, by sądzić, że tak. Po pierwsze, analizując to, czego uczy się głęboka sieć wyszkolona w rozpoznawaniu wzorców wzrokowych, widzimy, że poszczególne sztuczne neurony nabywają pewne właściwości detekcji cech, które obserwujemy w neuronach biologicznych. W niższych warstwach takiej sieci, czyli tych

położonych najbliżej obrazu z kamery, sztuczne neurony zazwyczaj rozwijają charakterystyki reakcji odpowiadające prostym cechom, takim jak wykrywanie plamek (jasnych plamek z ciemniejszym otoczeniem), krawędzi lub prostych wzorów geometrycznych, takich jak siatki i narożniki. Podobne reakcje zaobserwowano w komórkach zwojowych siatkówki w tylnej części oka oraz w pierwotnej korze wzrokowej, co zaobserwowano na przykład w eksperymentach Hubela i Wiesela. Sztuczne neurony na wyższych poziomach głębokiej sieci, innymi słowy te bliższe podejmowaniu decyzji klasyfikacyjnych, mają znacznie bardziej złożone właściwości detektora cech. Na przykład mogą reagować na złożone kształty, tekstury lub nieregularne wzory, na części obiektu, a nawet na całe obiekty. Podobnie, w miarę przesuwania się w górę szlaków wzrokowych mózgu, badacze odkryli neurony o coraz bardziej złożonych właściwościach reakcji. Inną cechą wspólną niektórych modeli głębokich sieci z mózgiem jest organizacja przypominająca piramidę. W tym przypadku największe macierze sztucznych neuronów znajdują się najbliżej obrazu, podczas gdy te najbliżej warstwy klasyfikacji często mają mniej neuronów. Oznacza to, że rozdzielczość na wyższych warstwach może być gorsza, ale może to również poprawić zdolność sieci do generalizacji, na przykład rozpoznawania obiektów w dowolnym miejscu obrazu (właściwość znana jako niezmienność przestrzenna). W miarę przesuwania się w górę szlaków wzrokowych mózgu następuje podobne zmniejszenie skali. Na przykład, zauważyliśmy wcześniej, że na każde dziesięć pręcików lub czopków przypada jedna komórka zwojowa siatkówki; Na każde siedem neuronów zwojowych siatkówki przypada tylko jeden neuron w obszarze V1 pierwotnej kory wzrokowej. Kolejnym pytaniem jest, czy mózg wykorzystuje coś podobnego do algorytmu uczenia się z propagacją wsteczną. Kiedy naukowcy początkowo się temu przyjrzeli, byli sceptyczni. Wydawało się, że mózgowi brakuje odpowiedniego rodzaju połączeń z projekcją wsteczną, umożliwiających propagację błędów. Niedawno pojawiły się dowody sugerujące, że w mózgu może zachodzić zjawisko zbliżone do propagacji wstecznej błędów.16 Co więcej, sukces w trenowaniu dużych modeli został osiągnięty również dzięki uczeniu się bez nadzoru i uczeniu się ze wzmocnieniem, z których niektóre są silnie inspirowane mechanizmami uczenia się występującymi w biologicznych obwodach neuronowych. Dzięki tym metodom nie ma również potrzeby etykietowania danych treningowych przez ludzi. Oczywiście sztuczne sieci neuronowe różnią się pod wieloma względami od obwodów neuronowych mózgu, a niektórzy badacze twierdzą, że różnice mogą być ważniejsze niż podobieństwa. Nie jestem pewien – na poziomie funkcjonalnym te głębokie modele sieci neuronowych zdają się w wielu kwestiach działać prawidłowo. AlexNet i podobne przełomy zapowiadały powszechny trend wykorzystywania technologii głębokich sieci w sztucznej inteligencji, szczególnie w wyzwaniach związanych z percepcją maszyn, które obejmują rozpoznawanie wzorców, osiągając poziom wydajności przekraczający możliwości człowieka w takich zadaniach, jak rozpoznawanie twarzy, klasyfikacja obrazów i obrazowanie medyczne. Sukces głębokich sieci sugeruje, że wczesna obietnica uczenia się w modelach sztucznej inteligencji przypominających mózgi jest spełniana, pomimo dłuższego niż oczekiwano oczekiwania i początkowego sceptycyzmu. Co uderzające, sieci te wydają się działać w sposób holistyczny, którego poszukiwał Dreyfus – rozwiązując szersze pytanie „Co to jest?", jednocześnie ustalając szczegóły. Ale czy istnieją obszary sztucznej inteligencji, w których technologie sieci neuronowych mogą nie odnieść sukcesu? Jednym z kandydatów może być język, jedna z umiejętności najbardziej odróżniających ludzi od zwierząt. Zbadamy to dalej.

MODELE JĘZYKÓW WIELOKROTNEGO UŻYTKU: PRZEŁOM W SZTUCZNEJ INTELIGENCJI?

Od samego początku jednym z kluczowych wyzwań dla SI było rozumienie i używanie języka ludzkiego. Jak zauważono , było to również wyzwanie, jakie postawił przed SI Alan Turing – czy maszyna mogłaby prowadzić rozmowę z człowiekiem, nie wiedząc, że rozmawia z artefaktem? Fakt, że żadne inne zwierzę nie rozwinęło tak zaawansowanej formy komunikacji jak język ludzki, również czyni to problemem szczególnym, który odwołuje się do naszych najgłębiej zakorzenionych intuicji dotyczących wyjątkowej

natury ludzkiej inteligencji. Pamiętacie test inteligencji WAIS i jego cztery podskale ? Były to rozumienie werbalne, pamięć robocza, organizacja percepcyjna i szybkość przetwarzania. Jak widzieliśmy, SI już teraz stawia przed ludźmi wyzwania w zakresie pamięci, percepcji i mocy obliczeniowej. Gdyby SI poradziła sobie z wyzwaniem językowym, pogląd, że maszyny nigdy nie dorównają ludzkiej inteligencji, straciłby jeden ze swoich głównych filarów. Istnieją pewne dowody na to, że przełom już nastąpił. W 2023 roku mogę prowadzić sensowną (dla mnie) rozmowę z modelem języka opartym na sztucznej sieci neuronowej. W istocie, pisząc te słowa, prowadzę poboczną rozmowę z ChatGPT 3.5, gdzie czerpię z niego pomoc w śledzeniu historii sztucznej inteligencji i rozważaniu kluczowych idei AI oraz ich wzajemnych powiązań. Oczywiście zdaję sobie sprawę, że ChatGPT może popełniać błędy; zauważyłem kilka z nich i dokładnie sprawdziłem w innych źródłach. Jednak w dużej mierze jest to trafne stwierdzenie. Jako sposób na szybkie uzyskanie użytecznych informacji, z pewnością wzbogaca mój repertuar narzędzi „rozszerzonego umysłu", obok wyszukiwarek, baz danych, e-papierów, e-booków i okazjonalnych drukowanych tomów z mojej biblioteczki. Zdolność ChatGPT i innych rozbudowanych modeli językowych do posługiwania się językiem zaskoczyła wiele osób, w tym mnie. Są powody, by być pod wrażeniem. W 2023 roku psycholog kliniczny Eka Roivainen, psycholog ewaluacyjny, który rutynowo testuje ludzką inteligencję, poddał ChatGPT test WAIS sprawdzający werbalny iloraz inteligencji.17 Nie było żadnego specjalnego traktowania – Roivainen wycinała i wklejała pytania z testu, a odpowiedzi zaznaczała tak, jak w przypadku uczestnika. Według Roivainen, ChatGPT „zdał" test celująco. Uśredniając wyniki pięciu testów werbalnego ilorazu inteligencji w skali WAIS, ChatGPT uzyskał werbalny iloraz inteligencji na poziomie 155 – lepszy niż 99,9% z 2450 osób, które zostały przetestowane w momencie standaryzacji skali. Jak sieci neuronowe stały się tak dobre w języku (przynajmniej według tego standardowego psychologicznego benchmarku)? Musimy cofnąć się do historii badań nad przetwarzaniem języka naturalnego w sztucznej inteligencji, aby się tego dowiedzieć.

CZY JĘZYK WYMAGA REGUŁ?

Wczesne postępy w dziedzinie sztucznej inteligencji sugerowały, że problem przetwarzania języka naturalnego wymaga rozwiązania opartego na regułach. Językoznawca i kognitywista Noam Chomsky słynnie krytykował behawiorystyczne teorie języka, takie jak te B. F. Skinnera, za to, że nie wyjaśniają one produktywności językowej, czyli zdolności osób posługujących się językiem do generowania zdań, które nigdy wcześniej nie zostały wypowiedziane. Według Chomsky'ego18, umysł musi posiadać złożone, wrodzone i specyficzne dla języka systemy przetwarzania, aby przyswajać i używać języka; to znaczy, że rodzimy się ze znajomością zbioru reguł gramatycznych, zwanego gramatyką uniwersalną, leżącego u podstaw wszystkich języków ludzkich. Naukowcy badający sztuczną inteligencję opartą na regułach szybko odrzucili również sieci neuronowe epoki Rosenblatta jako zbyt proste, by mogły być interesujące. Modele te opierały się wyłącznie na uczeniu się „skojarzeń", idei zapoczątkowanej przez Arystotelesa . Uczenie asocjacyjne leżało również u podstaw teorii behawiorystów, w tym oczywiście reguły uczenia się Rescorli-Wagnera. Oczywiście, choć uczenie się miało swoje miejsce w wyjaśnianiu, jak gołębie uczą się dziobać kształty na ekranach, wydawało się, że inteligentne reguły uczenia się nigdy nie wystarczą do rozwiązania problemu języka. Jak system, który uczy się na pamięć, mógłby kiedykolwiek stworzyć oryginalne zdanie? Język wydawał się również ewidentnie ograniczony regułami gramatycznymi. Weźmy na przykład zdanie: „Chłopiec, który gonił psa, potknął się o kota". Do czego odnosi się tu czasownik „potknął się"? Do psa, chłopca, czy nawet kota? Oczywiście, chodzi o chłopca (a także o kota), jak mogłoby powiedzieć wielu czterolatków. Według językoznawców, takich jak Chomsky, wiesz to, ponieważ twój mózg dekonstruuje to zdanie i odkrywa, że zawiera ono zdanie podrzędne „który gonił psa". Twoja wiedza o strukturze języka pozwala ci odłożyć to zdanie podrzędne na bok, podczas gdy przetwarzasz resztę, pozostawiając ci zrozumienie prostszej konstrukcji „chłopiec... potknął się o kota". Twój mózg zna również różne klasy słów. Na przykład wie, że czasowniki

takie jak „trip" wymagają podmiotu, który jest rzeczownikiem, w tym przypadku „chłopiec", i dopełnienia, które jest również rzeczownikiem, w tym przypadku „kot". Wreszcie, twój mózg wie również, że jeśli pościg miał miejsce w przeszłości („był"), to potknięcie musiało mieć miejsce również wtedy, stąd drugie użycie czasu przeszłego („tripped"). Wszystko to przemawia za systematyczną naturą języka i potrzebą, aby umysł znał reguły językowe, ale gdzie w sieci neuronowej mogłyby znajdować się reguły gramatyki uniwersalnej? Dla Jerry'ego Fodora, kolegi Chomsky'ego z Wydziału Mózgu i Nauk Kognitywnych MIT, problemem było to, że sieci neuronowe były subsymboliczne; Nie potrafili uchwycić idei symboli jako zastępstwa pojęć. Dla Fodora systematyczność i produktywność, które odnajdujemy w języku ludzkim, miały również kluczowe znaczenie dla innych dziedzin myślenia i rozumowania. Umysł, jak przypuszczał, musi używać własnego, prywatnego kodu symbolicznego do reprezentowania i manipulowania pojęciami oraz do wykonywania myślenia; nazwał ten wewnętrzny system symboli „językiem myśli". Na pierwszy rzut oka argumenty teoretyków języka wydawały się przekonujące. Nawet w złożonej sieci, takiej jak NetTalk, która potrafi nauczyć się czytać, istnieje jedynie las ważonych połączeń i nic, co przypominałoby reguły. Fakt, że NetTalk uzyskał zaledwie 95% w swoim zestawie testowym, pozostawił otwarte pytanie: czy różnica może wynikać ze znajomości przez ludzki umysł reguł wymowy? Debata między „symbolistami" (takimi jak Fodor) a koneksjonistami została dodatkowo ożywiona przez stronę koneksjonistów. Koneksjoniści argumentowali, że mózg używa miękkich ograniczeń zakodowanych przez wagi sieci, a nie sztywnych reguł. Co więcej, nie był to tylko „asocjacjonizm". Zdolność ukrytych neuronów do tworzenia nowych rodzajów reprezentacji wewnętrznych sprawiła, że modele koneksjonistyczne stały się „poznawcze" i skuteczniejsze niż wcześniejsze teorie uczenia się behawiorystów. W mózgu istniał pewien rodzaj wewnętrznego kodu, ale przybrał on formę masowo rozproszonych i stale zmieniających się wzorców aktywacji. W 1989 roku David Rumelhart, jeden z twórców metody backprop, napisał: „Krótko mówiąc, naszym celem jest zastąpienie metafory komputerowej metaforą mózgu". To była wojna o wpływy. Celem koneksjonistów było wyjaśnienie wszystkiego, co dotyczy umysłu, a język był kluczowym polem bitwy.

NAUKA JĘZYKA ZA POMOCĄ SIECI REKURENCYJNYCH

W latach 80. i na początku lat 90., kiedy debata między sztuczną inteligencją symboliczną a subsymboliczną była najbardziej zacięta, wciąż nie było jasne, kto lepiej rozumie, jak mózg przetwarza język. Trwały jednak ważne prace, które wskazywały na potencjał technologii sieci neuronowych, mimo że do powstania użytecznych narzędzi do przetwarzania języka sterowanych sieciami była jeszcze daleka droga. Jeden ważny zestaw wskazówek pochodził z badań nad rekurencyjnymi sieciami neuronowymi. Sieć rekurencyjna to taka, w której sygnał wyjściowy sieci (na pewnym poziomie) może płynąć wstecz i wpływać na swoje przyszłe dane wejściowe. Kontrastuje to z sieciami typu feed-forward najczęściej badanymi w sztucznej inteligencji, w tym z wieloma obecnie stosowanymi modelami głębokich sieci neuronowych, w których sygnały przepływają przez sieć tylko w jednym kierunku, od wejścia do wyjścia, i nigdy w drugą stronę (poza dostosowywaniem połączeń). Istnieje wiele powodów, dla których warto zainteresować się sieciami rekurencyjnymi. Jednym z nich jest fakt, że w biologicznych układach neuronowych występuje duża powtarzalność – nasze mózgi są pełne lokalnych sieci wysyłających sygnały do innych sieci i odbierających je z powrotem, często jednocześnie. Nawet w omówionych wcześniej szlakach percepcyjnych mózgu informacje przepływają w obu kierunkach, czego nie uwzględnia większość obecnych modeli głębokich sieci. Na przykład w układzie wzrokowym wyższe obszary kory mózgowej projektują z powrotem w kierunku niższych, w tym pierwotnej kory wzrokowej. Innym powodem do badania sieci rekurencyjnych jest ich przydatność w zrozumieniu przetwarzania zachodzącego w czasie. Aktywność mózgu nieuchronnie ma wymiar czasowy, ponieważ jej głównym wyzwaniem jest radzenie sobie w czasie rzeczywistym z ciągle zmieniającym się światem. W obwodach mózgowych generujących zachowania powtarzalność jest również częstym zjawiskiem, szczególnie w przypadku zachowań cyklicznych, takich jak chodzenie, pływanie czy oddychanie. W

istocie, każde miejsce w mózgu, które generuje wzorce aktywności, prawdopodobnie wykorzystuje powtarzalność. Ważnym przykładem procesu, który generuje wzorce aktywności w czasie, jest oczywiście język. Kiedy słuchasz lub mówisz, słyszysz lub generujesz słowa w sekwencji. Nawet czytając, będziesz czytać sekwencyjnie od początku do końca akapitu, nawet jeśli czytając szybko, możliwe będzie zinterpretowanie fragmentu tekstu większego niż jedno słowo na raz. W języku, podobnie jak w wielu innych procesach sekwencyjnych, to, co było wcześniej, jest ważne dla zrozumienia tego, co dzieje się teraz. Zatem, aby sieć neuronowa rozumiała język, musi być w stanie śledzić treść napotkaną wcześniej w interakcji; innymi słowy, musi mieć pamięć. Rekurencyjne sieci neuronowe mogą zapewnić formę pamięci, przekazując informacje z wyjścia sieci z powrotem do wejścia. Ten proces może „utrzymywać przy życiu" istotne informacje z wcześniejszej części interakcji bez konieczności przechowywania wszystkiego, co zostało powiedziane. Badacze zajmujący się sztuczną inteligencją opartą na regułach również doskonale zdawali sobie sprawę z potęgi rekurencji. Modele oparte na regułach w sztucznej inteligencji, w tym formalne modele języka naturalnego, w szerokim zakresie wykorzystują reguły rekurencyjne, czyli takie, które same się nazywają. Całe języki programowania sztucznej inteligencji, takie jak Lisp i Prolog, zostały specjalnie zaprojektowane, aby ułatwić wykorzystanie potęgi rekurencji. W drugiej fali koneksjonizmu, w latach 80. XX wieku, zaczęto już badać potencjał rekurencyjnych sieci neuronowych. Ważnym pytaniem było sprawdzenie, czy w praktyce możliwe jest wykorzystanie sieci rekurencyjnych do generowania zachowań sekwencyjnych i pełnienia funkcji pamięci krótkotrwałej. Jeśli tak, czy można je zastosować w ostatecznym wyzwaniu, jakim jest nauka języka? Na początku lat 90-tych Jeffry Elman dołączył do Wydziału Kognitywistyki w San Diego, aby podjąć się tego wyzwania. Jednym z problemów, które postanowił zbadać, była wysoce uproszczona wersja języka angielskiego. Dokładniej rzecz ujmując, Elman postanowił wytrenować prostą sieć rekurencyjną, składającą się z nieco ponad dwustu sztucznych neuronów i nieco poniżej trzech tysięcy regulowanych połączeń, na dużej liczbie zdań składających się z od trzech do szesnastu słów, ale wszystkie skonstruowane przy użyciu bardzo ograniczonego słownictwa, liczącego zaledwie dwadzieścia trzy słowa. Zadaniem sieci Elmana było przewidywanie kolejnego słowa w zdaniu, a sieć została wytrenowana, aby poprawić tę umiejętność, wykorzystując mechanizm backprop. Innymi słowy, za każdym razem, gdy sieć przewidziała błędne słowo, występował błąd zwrotny, a algorytm uczenia wprowadzał niewielkie zmiany we wszystkich siłach połączeń w sieci. Hipoteza Elmana zakładała, że jeśli sieć była w stanie nauczyć się tego zadania, oznaczałoby to, że zinternalizowała coś ważnego w strukturze języka. Sieć Elmana była w stanie nauczyć się zadania predykcyjnego, które postawił, jak pokazano na rysunku 5.4. Należy zauważyć, że sieć ma warstwy pośrednie, które tworzą rozproszone reprezentacje słów w zdaniu, z zestawem jednostek „kontekstowych", które przetwarzają dane wyjściowe jednej z warstw pośrednich, aby stały się częścią jej własnych danych wejściowych w kolejnym kroku czasowym. Warstwy wejściowa i wyjściowa używają dokładnie tej samej reprezentacji, z jedną jednostką reprezentującą każde z dwudziestu trzech słów (i kilkoma dodatkowymi jednostkami dla niesłów) oraz ze stałą warstwą kodowania/dekodowania na każdym końcu, która konwertuje to na/z rozproszonego wzorca aktywacji. Po treningu Elman przedstawił sieci każde możliwe słowo wejściowe i zarejestrował wzorzec aktywacji utworzony w warstwie pośredniej. Porównując te wzorce dla różnych słów, był w stanie wykazać, że jego sieć nauczyła się grupować pewne typy słów, które odgrywają podobne role w zdaniach. Na przykład rzeczowniki miały podobieństwa w sposobie, w jaki były reprezentowane w sieci, podczas gdy czasowniki również tworzyły własne klastry. Co więcej, różne typy rzeczowników miały tendencję do grupowania się. Na przykład rzeczowniki, takie jak imiona ludzkie lub nazwy zwierząt, tworzyły klastry, ponieważ często określają one aktora, czyli „podmiot" w zdaniu. Jednocześnie grupowały się rzeczowniki odnoszące się do przedmiotów nieożywionych – te rzeczowniki odnoszą się do przedmiotów, na które często działa się, pełniąc funkcję „obiektu" w zdaniu. Elman odkrył również, że jego sieć była w stanie zakodować pewne dalekosiężne zależności gramatyczne. Na przykład w zdaniu „Dziewczyna, która goniła psa, który ugryzł chłopca, lubi lody" sieć

była w stanie poprawnie przewidzieć czasownik w liczbie pojedynczej „lubi", który pasowałby do podmiotu w liczbie pojedynczej „dziewczyna", który pojawił się dziewięć słów wcześniej. Jest to coś, z czym teoretycy symboliczni twierdzili, że modele koneksjonistyczne będą miały trudności. Zauważmy, że sieć Elmana została wyszkolona jedynie do przewidywania następnego słowa; nie otrzymała żadnego szkolenia dotyczącego typów gramatycznych, takich jak rzeczowniki i czasowniki czy podmioty i dopełnienia. Ucząc się przewidywać, sieć niejawnie nabyła wiedzę, którą można powiązać z tymi pojęciami gramatycznymi. Zauważmy również, że sieć nie nauczyła się dosłownie reguł gramatyki; zamiast tego nauczyła się setek wag połączeń, których zachowanie można w przybliżeniu opisać za pomocą gramatyki formalnej. Dla innego koneksjonisty, Paula Smolensky'ego, ten subsymboliczny poziom reprezentacji pozostaje w relacji do teorii opartych na regułach Chomsky'ego i innych, takich jak relacja między mechaniką kwantową a klasycznymi prawami ruchu Newtona w fizyce. Fizyka Newtona dostarcza użytecznego, choć niedokładnego opisu tego, jak ruch zachodzi w świecie, ale mechanika kwantowa daje głębszy i dokładniejszy obraz. Dla Smolensky'ego teorie oparte na regułach mogą działać jako przybliżenie tego, co działo się w mózgu, ale aby w pełni zrozumieć, jak z mózgu wyłania się inteligencja, trzeba zejść o poziom niżej.

DUŻE MODELE JĘZYKOWE

Sieć rekurencyjna Elmana, z zaledwie trzema tysiącami regulowanych wag, pokazała, że w zasadzie możliwe jest odkrycie struktury gramatyki poprzez naukę przewidywania następnego słowa w zdaniu. Co ciekawe, duże modele językowe (LLM), takie jak ChatGPT firmy Open AI, PaLM firmy Google i LLama firmy Meta, wykorzystują w dużej mierze tę samą sztuczkę, ale tym razem z całą mocą i skalą nowoczesnego głębokiego uczenia. Na przykład ChatGPT 3.5 ma 175 miliardów parametrów, około sześćdziesiąt milionów razy więcej niż sieć Elmana i około trzy tysiące razy więcej niż AlexNet. Modele LLM należą do szerszej klasy generatywnych modeli sztucznej inteligencji, które uczą się generować nowe dane, które mogłyby należeć do zbioru danych, na którym zostały wytrenowane. W przypadku modeli LLM, modele te są trenowane na ogromnych bazach danych tekstu generowanego przez ludzi, zebranych z internetu. Jak już zauważyliśmy, zdolność LLM do generowania nowego tekstu oznacza, że sztuczna inteligencja może teraz twierdzić, że osiągnęła poziom ludzki w testach inteligencji werbalnej. Wykorzystamy aktualną, darmową wersję LLM OpenAI, ChatGPT 3.5, aby zbadać, jak działają te potężne sztuczne inteligencje. Na początku lat 2010. naukowcy eksperymentowali z coraz większymi sieciami rekurencyjnymi do przetwarzania języka, wykorzystując nienadzorowaną metodę treningu Elmana, w której celem sieci jest przewidywanie kolejnego słowa. Podstawową wartością rekurencyjności w języku jest zdolność do zachowania śladu istotnych wcześniejszych fragmentów tekstu – formy pamięci roboczej i selektywnej uwagi. W 2015 roku Ashish Vaswani i zespół badaczy z Google odkryli, że podobne wyniki można osiągnąć, zachowując przeszłe elementy sekwencji jako dane wejściowe dla sieci i zamiast rekurencyjności stosując w pełni wyprzedzający mechanizm uwagi, znany jako transformator. Transformator działa jak głowica odtwarzająca w magnetofonie taśmowym, wybierając jedną część taśmy – w tym przypadku sekwencję słów – do skupienia się w dowolnym momencie. Wielogłowicowy model transformatora pozwala na skupienie uwagi na wielu wcześniejszych częściach sekwencji, z przełączaniem uwagi w razie potrzeby w trakcie przetwarzania tekstu. Modele transformatorów są kluczowe dla działania ChatGPT; w istocie, część GPT w nazwie ChatGPT oznacza „Generative Pre-trained Transformer" (Generatywny Wstępnie Wyszkolony Transformator). Kiedy ChatGPT przetwarza tekst, bierze sekwencję słów i przekształca ją we wzorzec aktywacji obejmujący tysiące sztucznych neuronów, kodujących obecne słowa i ich pozycje w sekwencji. Ściśle rzecz biorąc, model działa z częściami słów, zwanymi tokenami, poprzez dekonstrukcję słów na podsłowa, które mogą być informacyjne i pomagać sieci w generalizacji. Na przykład „kolorowy" można tokenizować jako „kolor" i „pełny", a „wędrował" jako „wędrował" i „ed". W ChatGPT 3.5 wzorzec wejściowy może zakodować do 4096 sekwencyjnych tokenów (wersja 4

pamięta do 32 000 tokenów). Dane wejściowe są przepuszczane przez dziewięćdziesiąt sześć warstw transformatorowych, z których każda składa się z połączenia głowic uwagi i sieci sprzężenia zwrotnego. Model generuje następnie dane wyjściowe, które są predykcją możliwych kolejnych tokenów. Dane te są porównywane z rzeczywistym kolejnym tokenem w strumieniu tekstu i, podobnie jak w sieci Elmana, różnica ta jest wykorzystywana do trenowania warstw transformatorowych za pomocą propagacji wstecznej błędów. Innymi słowy, podobnie jak w wielu sieciach neuronowych, uogólniona reguła delty jest wykorzystywana do przeprowadzenia spadku gradientu w przestrzeni parametrów sieci. Należy zauważyć, że ChatGPT, podobnie jak inne współczesne modele LLM, nie tylko przewiduje kolejny token. Zamiast tego, przypisuje prawdopodobieństwo wszystkim możliwym tokenom. W przypadku generowania języka, pozostawia to pewien mechanizm do decydowania, który token zostanie użyty w bieżącym zadaniu. Kiedy Alec Radford i współpracownicy z OpenAI rozpoczęli prace nad serią modeli językowych ChatGPT, odkryli, że tego rodzaju sieć, po wytrenowaniu, może być użytecznie łączona z systemami uczenia nadzorowanego lub wzmacniającego w celu precyzyjnego dostrojenia lub realizacji konkretnych zadań językowych. Na przykład, OpenAI wytrenowało rozszerzenia swojego modelu, które mogły streszczać tekst, przeprowadzać rozumowanie oparte na zdrowym rozsądku i klasyfikować zdania (na przykład jako gramatyczne lub niegramatyczne). Jedną z takich rozszerzonych wersji ChatGPT jest internetowy chatbot OpenAI, który pobiera serię podpowiedzi od użytkownika i za każdym razem generuje odpowiedź tekstową. W tym przypadku dodatkowy mechanizm wybiera konkretny token z rozkładów prawdopodobieństwa możliwych tokenów generowanych przez model główny. Wybrany token staje się następnie częścią danych wejściowych modelu, które są wykorzystywane do generowania kolejnych tokenów. Należy zauważyć, że poprzez włączenie własnego wyjścia do zmienionej sekwencji danych wejściowych, model implementuje formę rekurencji. Korzystanie z ChatGPT pozwala chatbotowi tworzyć zdania i akapity, które nigdy wcześniej nie zostały napisane. Aby uniknąć zbytniej przewidywalności, część chatbota generująca tekst może wprowadzić umiarkowaną ilość szumu do procesu wyboru tokenów. Jest to parametr, który można zmieniać, ale eksperymenty wykazały, że prawdopodobieństwo około 0,8 na wybranie najbardziej prawdopodobnego tokenu (innymi słowy, jedna na pięć szans na wybranie mniej prawdopodobnego tokenu) zapewnia dobrą równowagę między trzymaniem się tematu a nadmiernym powtarzaniem słów lub idei. Przyjrzyjmy się wynikom z ChatGPT 3.5. Wprowadziłem następującą instrukcję do nowej sesji: „Napisz sonet w stylu Williama Szekspira o miłości skrzypka do jego psa-robota”. Oto, co otrzymałem:

Na scenie, gdzie cienie delikatnie tańczą,

Palce skrzypka snują czuły czar,

Jego melodię, szeptany słodki romans,

Z brzęczącymi strunami, opowieść o miłości do opowiedzenia.

Obok niego stoi istota wykuta ze stali,

Wierny pies, z obwodami i przekładniami,

Jednak w jego oczach ciepło, które wydaje się tak prawdziwe,

Lojalne serce, nieskalane śmiertelnymi lękami.

Razem tworzą rzadki sonet,

Westchnienia skrzypiec, cichy szum robota,

Duet grany z pasją i troską,

Jak dwie dusze w idealnej harmonii.

W krzemie i strunach miłość odnajduje swoją drogę,

Skrzypek i jego pies-robot podczas zabawy.

Zasady pisania sonetów, które są dość surowe, są tu ściśle przestrzegane, w tym schemat rymów ABAB CDCD EFEF GG. To, czy ten wiersz Ci się podoba, czy nie, to kwestia gustu. Jestem pewien, że niektórzy czytelnicy uznają go za patetyczny. Ja uznałem go za dość wzruszający, choć nieco banalny. Niezależnie jednak od tego, czy uznasz go za poezję, czy (robotyczny) wierszyk, mam nadzieję, że nadal będziesz w stanie docenić to osiągnięcie. Jest to prawie na pewno pierwszy wiersz o miłości między skrzypkiem a psem-robotem, więc jest tu pewna oryginalność – a może nawet kreatywność. ChatGPT nie zna żadnych wyraźnych reguł językowych i dowiedział się o poezji, rymie, miłości, skrzypkach i robotach jedynie z istniejących przykładów języka ludzkiego. Jednak zwracanie uwagi na innych jest również w dużej mierze sposobem, w jaki ludzie uczą się swojego ojczystego języka. Chomsky i Fodor argumentowali, że do opanowania języka potrzebna jest znaczna wrodzona wiedza specyficzna dla danego języka, w tym pewna gramatyka uniwersalna. Mamy tu system niemal bez znajomości języka, a już na pewno bez gramatyki, który nauczył się tworzyć oryginalne fragmenty tekstu, które są uporządkowane, gramatyczne i sensowne (dla nas, ludzi). Ale czy generatywne sztuczne inteligencje, takie jak ChatGPT, to po prostu bardzo rozbudowane programy, które naśladują aspekty ludzkiej inteligencji, ale wciąż brakuje im kluczowych elementów, by jej dorównać lub ją przewyższyć?

W KIERUNKU SZTUCZNEJ INTELIGENCJI OGÓLNEJ

Wydajność głębokiego uczenia i generatywnych SI jest imponująca i stanowi przełom w rozwiązywaniu niektórych z najtrudniejszych problemów SI. Jednak niektórzy nie są przekonani, że to prawdziwa inteligencja. Zbadamy tę kwestię szerzej w tym rozdziale i omówimy, co jeszcze SI może potrzebować, aby dorównać inteligencji ludzkiej lub ją przewyższyć. Zastanowimy się również nad tym, co niektórzy badacze SI opisują jako budowanie Sztucznej Inteligencji Ogólnej, czyli AGI – formy SI, która ma być równie elastyczna jak inteligencja ludzka, jeśli chodzi o zdolność do działania w różnych dziedzinach i adaptacji do nowych wyzwań.

CZY DUŻE AI'S UCZĄ SIĘ CZEGOŚ? CIEKAWE?

Po pierwsze, rozważmy jedną z krytyk obecnej sztucznej inteligencji, w postaci sztucznej inteligencji generatywnej, a w szczególności dużych modeli językowych (LLM). Sugeruje to, że modele te nie są tak inteligentne, jak się wydaje i wykorzystują moc obliczeniową i ogromną pojemność pamięci, aby sprawiać wrażenie inteligentnych. Krytykę tę wysunął pisarz, psycholog i niekiedy sceptyk sztucznej inteligencji Gary Marcus, który napisał w 2022 roku: „W przypadku GPT-3 mechanizm predykcji jest zasadniczo regurgitacją; takie systemy są trenowane na dosłownie miliardach słów cyfrowego tekstu; ich darem jest znajdowanie pasujących elementów, na których zostały wytrenowane".1 Ten rodzaj dopasowywania wzorców przypomina to, co Dreyfus krytykował w latach 70. XX wieku jako dopasowywanie pod kątem „podobieństwa fizycznego", które uważał za niewystarczające do wyjaśnienia ludzkiej inteligencji (patrz rozdział 4). Można by sobie wyobrazić, biorąc pod uwagę ogromny rozmiar tych modeli, że tego rodzaju gloryfikowane „wyszukiwanie w tabeli" jest możliwe. Na przykład ChatGPT 3.5 ma 175 miliardów parametrów. Czy może to po prostu przechowywać wiele użytecznych fragmentów tekstu, które są następnie łączone w celu uzyskania wiarygodnych odpowiedzi? W końcu tak właśnie działały wczesne chatboty, dopasowując wzorce do słów docelowych w podpowiedzi, a następnie wykorzystując ten konkretny wzorzec do wyszukiwania odpowiednich odpowiedzi w dużej tabeli. Jeden z pierwszych chatbotów, „Eliza", stworzony przez Josepha Weizembauma w połowie lat 60. XX wieku, który słynął z tego, że czasami oszukiwał niektórych

ludzi, działał dokładnie w ten sposób. Chociaż prawdą jest, że ChatGPT to niezwykle potężna forma pamięci, odpowiedź brzmi, że nic tak prostego, jak proste dopasowywanie wzorców, nie mogłoby tu działać. Rozpoznawanie i dopasowywanie wzorców z pewnością istnieje, ale jest ono znacznie bardziej abstrakcyjne i subtelne niż jakiekolwiek proste „odświeżanie" zapamiętanych ciągów słów. Jednym z powodów, dla których proste wyszukiwanie w pamięci nie zadziała, jest to, że wiele słów ma wiele znaczeń (polisemia). Oznacza to, że możemy zrozumieć poszczególne słowa tylko poprzez rozważenie kontekstu innych słów, z którymi się pojawiają. Jak widzieliśmy , sieć Elmana nie realizowała prostego dopasowywania wzorców. Zamiast tego, był w stanie znaleźć strukturę gramatyczną w zdaniach, na przykład śledząc dalekosiężne zależności między słowami (choć ucząc się zaledwie ułamka języka ludzkiego). Biegłość ChatGPT w posługiwaniu się językiem, w tym jego zdolność do interpretowania słów w kontekście, mówi nam, że nie może on również dokonywać prostego dopasowywania wzorców. Zamiast tego, abstrahuje od tekstu dosłownego, aby znaleźć głęboką strukturę, która reprezentuje sposób, w jaki słowa łączą się, aby wyrazić idee (zbadamy to szerzej w odniesieniu do innych SI w dalszej części). Elman był w stanie wykazać, że jego sieć uczy się ukrytej struktury języka. ChatGPT robi to samo w szerszej perspektywie. Zadanie zrozumienia, jak ChatGPT rozwiązuje problemy językowe, dopiero się zaczyna – nie mamy jeszcze teorii języka – ale mamy dowód istnienia na możliwość sztucznego użycia języka, który wykracza daleko poza dopasowywanie wzorców na podstawie powierzchownego podobieństwa. Innym powodem, dla którego wątpimy, czy jakiekolwiek podejście siłowe przyniesie skutek, jest ogromna przestrzeń poszukiwań języka. Matematyk Steven Wolfram omawia pomysł prostego przechowywania informacji o częstotliwości występowania słów w sekwencjach o różnej długości, zwanych n-gramami. Wolfram szacuje, że w języku angielskim w powszechnym użyciu jest około czterdziestu tysięcy słów; jednak jeśli spojrzymy na pary słów, czyli 2-gramy, dla których pełny model języka mógłby przechowywać statystyki, jest ich już około 1,5 miliarda. Jeśli przejdziemy do 3-gramów, jest ich prawie sześćdziesiąt bilionów. W miarę zbliżania się do długości typowego zdania (piętnaście do dwudziestu słów), liczba możliwych kombinacji staje się astronomiczna. Pamiętacie, jak w rozdziale 4 widzieliśmy, że przestrzeń poszukiwań dla szachów ma więcej stanów niż atomów w znanym wszechświecie? To samo dotyczy języka. Wreszcie, jeśli porównamy ChatGPT do ludzkiego mózgu, to jego rozmiar – 175 miliardów parametrów – wygląda skromnie. Stanowi to mniej niż jedną piątą 1% całkowitego szacowanego rozmiaru ludzkiego konektomu (sto bilionów połączeń). Tylko w przypadku obszaru Broki, o którym wiadomo, że odpowiada za przetwarzanie języka, w tym gramatyki, ostrożne szacunki wskazują na 1,5 biliona połączeń, czyli prawie dziesięciokrotnie więcej niż ChatGPT 3.5. Jeśli uważasz, że głębokie SI wykorzystują ogromną pojemność pamięci i moc obliczeniową do rozwiązywania problemów językowych, to po prostu podążają za przykładem natury.

CZY SZTUCZNA INTELIGENCJA ROZUMIE?

Powszechną krytyką systemów sztucznej inteligencji jest to, że przetwarzają one jedynie symbole za pomocą reguł i nie rozumieją, co cokolwiek oznacza (przypomnijmy sobie eksperyment myślowy Searle'a „Chiński pokój"). Czy zatem systemy LLM, takie jak ChatGPT, oddają znaczenie słów? Odpowiedź brzmi: i tak, i nie. Filozofowie języka, począwszy od Gottloba Fregego, którego spotkaliśmy jako jednego z twórców logiki, rozróżniają między znaczeniem słowa a jego desygnatem. Sens odnosi się do mentalnej treści słowa, w tym jego relacji do innych słów. Na przykład słowa często występują w grupach – na przykład psy, koty, króliki i złote rybki są częścią grupy, którą moglibyśmy nazwać zwierzętami domowymi. Słowa zyskują również swój sens poprzez relację ze słowami, do których są podobne (synonimy) lub których przeciwieństwa (antonimy). Sens słowa zależy również od szerszego kontekstu, w którym się pojawia – fragment tekstu otaczający słowo jest często istotny dla ujednoznacznienia jego znaczenia. ChatGPT i podobni LLM-owie zdobywają dużą wiedzę na temat relacji między słowami, co potwierdza ich zdolność do odpowiadania na pytania dotyczące kategorii semantycznych, synonimów/antonimów itd. oraz do prawidłowej interpretacji słów wieloznacznych w

kontekście. W tym zakresie LLM-owie wyraźnie wiedzą wiele o znaczeniu w odniesieniu do sensu. Z drugiej strony, desygnat, według Fregego, opisuje obiekt lub obiekty w świecie, do którego odnosi się słowo lub wyrażenie. Na przykład zdanie „The Arctic Monkeys are a rock band from Sheffield" opisuje prawdziwy zespół rockowy, The Arctic Monkeys, i miasto, z którego pochodzi, Sheffield w Anglii. ChatGPT może wiele powiedzieć o Sheffield, w tym o jego położeniu geograficznym, historii i kulturze, a także o zespole Arctic Monkeys, w tym o tym, że powstał w Sheffield w 2002 roku. Jednak wszystko to ma związek ze znaczeniem – relacjami tych słów do innych słów – ChatGPT w ogóle nie ma wiedzy o desygnatach. Jeśli nigdy nie słyszałeś muzyki Arctic Monkeys, możesz jedynie twierdzić, że o nich wiesz; nie znałbyś ich tak, jak zna ich słuchacz muzyki. Jeśli nigdy nie byłeś w Sheffield, to tak naprawdę nie znasz Sheffield (więc powinieneś je odwiedzić!). ChatGPT jest w podobnej sytuacji, tyle że o wiele gorszej. Student LLM nigdy nie słyszał żadnej muzyki, a co dopiero indie-rocka czy post-punku, które pojawiały się w północnej Anglii na początku XXI wieku. Mówiąc ogólnie, ChatGPT nie ma bezpośredniego doświadczenia z czymkolwiek. Wszystko, co model wie o znaczeniu słów, dotyczy po prostu tego, jak słowa odnoszą się do innych słów. Zaskakujące w przypadku LLM jest to, jak daleko można dojść, opierając się wyłącznie na sensie; to naprawdę zdumiewające, że ChatGPT potrafi tak płynnie omawiać złożone idee bez odniesienia. Skupiliśmy się tutaj na LLM, ale postęp dokonuje się również w innych dziedzinach, na przykład w generowaniu obrazów, gdzie modele takie jak Midjourney i Dall-E potrafią generować obrazy na podstawie podpowiedzi tekstowych (co zostanie omówione dalej). Modele te ponownie wykorzystują możliwość pozyskiwania dużych ilości danych treningowych z internetu; w tym przypadku obrazów, które zostały opatrzone podpisami od widzów. W tego rodzaju sztucznej inteligencji system bazowy uczy się o obrazach i tekście razem, w ramach jednego modelu. Modele te kodują relację między językiem a zupełnie inną dziedziną, obrazami wizualnymi, co pokazuje jeden ze sposobów, w jaki słowa mogą nabierać znaczenia. Na przykład sztuczna inteligencja generująca obrazy, taka jak Dall-E, zna słowo „krzesło", ale może również generować obrazy przedstawiające wygląd krzeseł. Może to nie być odniesienie w pełnym tego słowa znaczeniu – obraz krzesła nadal nie jest krzesłem – ale obecnie istnieje więcej elementów krzesła niż wcześniej. Ucząc się tego rodzaju reprezentacji łączonych, SI mogą jawnie łączyć znaczenie słów z wieloma rodzajami danych pozajęzykowych. SI generujące muzykę potrafią teraz słuchać utworu i generować nową muzykę w podobnym stylu. Połącz tę SI generującą muzykę ze SI języka, a połączony system będzie mógł dowiedzieć się czegoś więcej o zespole Arctic Monkeys i ich muzyce. Taka SI mogłaby na przykład powiedzieć coś pożytecznego o muzycznym podobieństwie między Arctic Monkeys a innymi zespołami postpunkowymi. Połącz język, wideo i dźwięk w jednej sztucznej inteligencji, a otrzymasz inteligentny, multimodalny system językowo-pamięciowy, który jest bardziej podobny do tego w twojej głowie. Przypomnij sobie korowe obszary asocjacyjne. Wiele z nich jest silnie multimodalnych. Na przykład obszar kory ciemieniowej, zwany LIP, łączy sygnały wzrokowe, słuchowe, somatyczne i sensoryczne oraz przedsionkowe (ucho wewnętrzne), umieszczając je we wspólnym układzie odniesienia, który może być używany do kierowania ruchem w otoczeniu.5 Najnowsze osiągnięcia w dziedzinie sztucznej inteligencji, takie jak multimodalne modele generowania obrazu, zaczynają pokazywać, jak to osiągnąć. Czy to jednak znaczenie? Filozof Stevan Harnad, opierając się na myślowym eksperymencie Johna Searle'a „Chiński pokój", argumentowałby, że reprezentacje we wszystkich tych generatywnych SI wciąż nie są ugruntowane. Powodem jest to, że interpretacja ich wyników — obrazów, tekstu, dźwięku itd. — leży ostatecznie w umyśle obserwatora (użytkownika SI), a nie w maszynie. W końcu, na jednym poziomie, te SI po prostu manipulują ogromnymi tablicami liczb zgodnie z bardzo złożonymi regułami. Wygląda to bardzo podobnie do Searle'a w chińskim pokoju manipulującego bezsensownymi zawijasami. Czy jest tu cokolwiek, co byłoby równoważne ludzkiemu znaczeniu? Jak omówiono na początku tej książki, ludzie są inteligentni, a nie ich mózgi. Mózgi są pełne rozległych wzorców aktywności elektrycznej, a komórki mózgowe komunikują się za pomocą impulsów elektrycznych i chemii, więc trudno jest dostrzec tam również znaczenie, zwłaszcza jeśli po prostu zajrzysz do środka.

Dla wielu filozofów języka, podążających za Fregem, znaczenie ma charakter relacyjny – pojawia się, gdy rozważa się osobę i jej interakcję ze światem i innymi ludźmi. To prowadzi nas do kolejnego tematu, znaczenia działania w świecie dla ugruntowania znaczenia w sztucznej inteligencji.

CIAŁA I ROZSZERZONE UMYSŁY

Czy istnieje sposób, aby sztuczna inteligencja (AI) rzeczywiście rozumiała świat, a język, którym się posługuje, miał odniesienia w taki sam sposób, jak język ludzki? Tak, prawdopodobnie tak, i polega on na nadaniu sztucznej inteligencji ciała w postaci robota. Inteligencja w organizmach biologicznych występuje w ucieleśnionych systemach, w których osadzony układ nerwowy oddziałuje ze światem za pośrednictwem ciała. Chociaż ludzie stanowią dowód istnienia ucieleśnionej inteligencji ogólnej, nie istnieje naturalny model tego, jak mogłaby wyglądać całkowicie odcieleśniona inteligencja. Dlatego nie wiemy, czy inteligencja bez ciała jest w ogóle możliwa. Ucieleśnienie jest ważne dla ludzkiej inteligencji z kilku powodów. Po pierwsze, większość czynności mózgu wiąże się ze ścisłym powiązaniem percepcji z działaniem w celu wykonywania odpowiednich zachowań w świecie, które są zorganizowane czasowo i przestrzennie oraz ukierunkowane na cel. To samo dotyczy inteligencji w szerszym znaczeniu. Nasze ciała również ułatwiają mózgowi rozwiązywanie niektórych problemów. Na przykład, nasza zdolność jako zwierząt dwunożnych (dwunożnych) do chodzenia i biegania po nierównych powierzchniach zależy w równym stopniu od mechanicznej budowy naszych nóg i stóp, co od obwodów nerwowych kontrolujących ich ruch. Na przykład, łuki stopy, wraz z podeszwą, ścięgnami i więzadłami, zapewniają naturalną amortyzację, która pozwala na pochłanianie energii w kontakcie z podłożem, a następnie jej uwalnianie, gdy stopa się odrywa. Jednocześnie elastyczność palców i staw skokowy odgrywają ważną rolę w utrzymaniu równowagi. Elegancka i misterna konstrukcja ludzkiej stopy, opisana przez Leonarda da Vinci jako „arcydzieło inżynierii i dzieło sztuki", przyczynia się do wyjątkowej stabilności i efektywności energetycznej ludzkiego chodzenia i biegania, która nie ma sobie równych wśród współczesnych robotów. Ludzie, jako gatunek technologiczny, są również niezwykle biegli w eksploatowaniu świata i kształtowaniu go jako rozszerzenia naszej inteligencji. Na przykład, jak wspomniano wcześniej, filozof Andy Clark omawiał, jak przekazujemy część pracy umysłowej zewnętrznym narzędziom, takim jak długopis i papier, kalkulator czy komputer. Integrujemy również urządzenia zewnętrzne, takie jak protezy kończyn, instrumenty muzyczne, samochody i rowery, jako rozszerzenia naszych ciał. Język stanowi dodatkową formę rozszerzenia poznawczego, która jest zarówno wewnętrzna (mowa wewnętrzna), jak i zewnętrzna (dzielona z innymi), a tym samym służy do wspierania i przyspieszania naszego myślenia i rozwiązywania problemów.

CAŁKOWITY TEST TURINGA

Ucieleśnienie i koordynacja percepcji z działaniem są tak ważne dla ludzkiej inteligencji, że zaproponowano alternatywę dla testu Turinga, w której wyzwaniem jest nie tylko dorównanie ludzkim możliwościom językowym w odcieleśnionym środowisku konwersacyjnym, ale także odtworzenie naszych zdolności robotycznych. Innymi słowy, oprócz zdolności do komunikowania się jak człowiek, aby zdać „całkowity test Turinga" zaproponowany przez Stevana Harnada, sztuczna inteligencja powinna również być w stanie działać w świecie tak, jak człowiek. Chociaż nie ma realnej perspektywy, aby robot zdał całkowity test Turinga, dziedzina robotyki poczyniła znaczące postępy w tworzeniu form ucieleśnionej inteligencji podobnej do ludzkiej. Roboty potrafią teraz chodzić, a czasem biegać na dwóch nogach. Potrafią tworzyć mapy i nawigować w otoczeniu, chwytać, przenosić i manipulować przedmiotami. Obecnym robotom brakuje zwinności, zręczności i wytrzymałości ludzi, ale ta luka się zmniejsza. Naukowcy z dziedziny robotyki odkryli, że warstwowe architektury sterowania, łączące w sobie szybkie, ale nieelastyczne, oraz wolniejsze, ale bardziej przemyślane procesy, mogą najlepiej sprawdzać się w tworzeniu inteligencji robotów. Innymi słowy, architektury te zaczynają przypominać te, które spotykamy w ludzkim mózgu . Systemy robotyczne wymagają również większej liczby

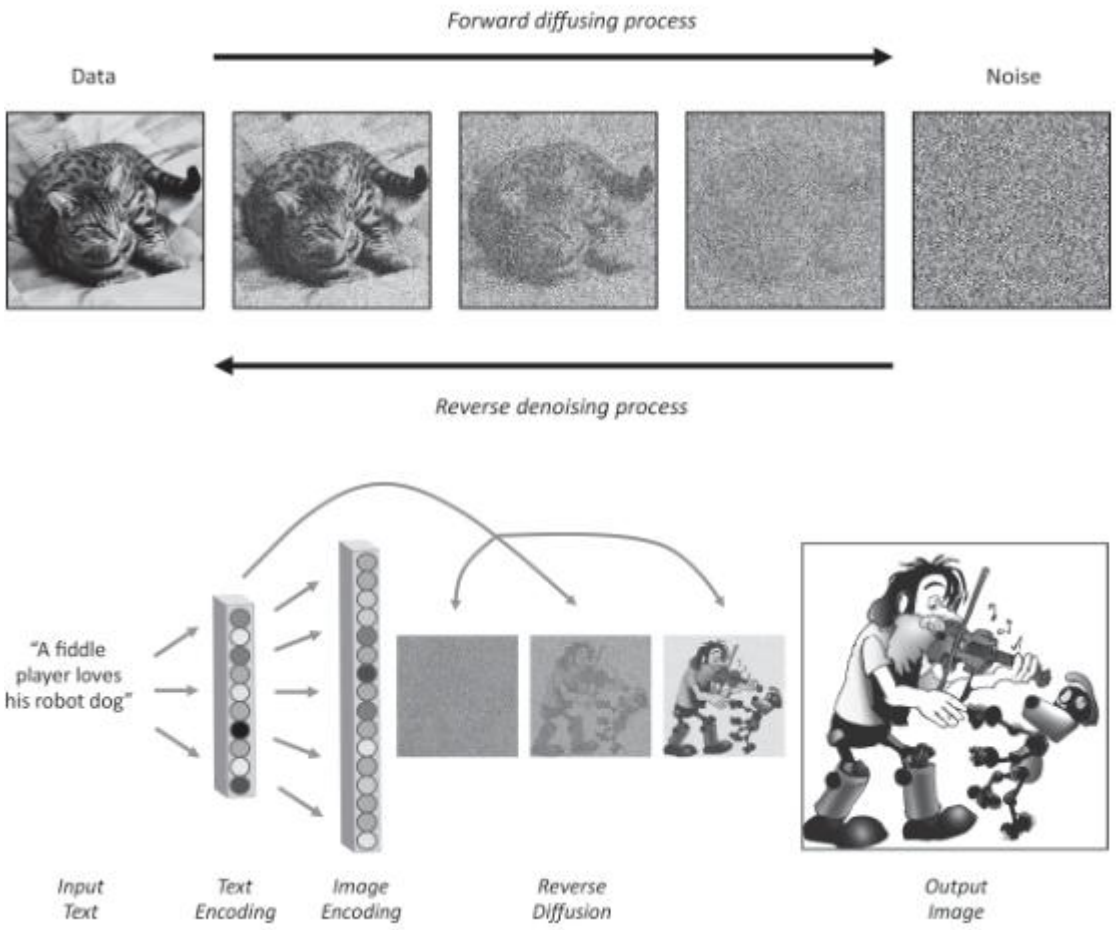
podstawowych elementów sterowania, przy czym sterowanie predykcyjne zyskuje na znaczeniu, co omówiono poniżej. Czy ucieleśniona sztuczna inteligencja, która zdałaby ten totalny test Turinga, rozumiałaby świat w pełnym tego słowa znaczeniu? Robot osadzony w środowisku i wchodzący z nim w interakcję z pewnością powinien być w stanie lepiej zrozumieć desygnaty języka niż bezcielesna sztuczna inteligencja. Na przykład, dla robota dwunożnego krzesło, oprócz tego, że jest czymś, co potrafi rozpoznać i nazwać, byłoby obiektem fizycznym, który mógłby przesuwać, podnosić, przewracać, na którym mógłby usiąść lub z którego mógłby spaść. Innymi słowy, krzesło miałoby wiele możliwości działania dla robota dwunożnego, tak jak dla człowieka. Sukces w teście Turinga powinien być również bardziej odporny na filozoficzne wyzwanie, jakim jest Chiński Pokój. Jak pisze Harnad:

Argument Searle'a logicznie zawodzi w przypadku wersji testu Turinga dotyczącej robota, ponieważ symulując go [robota] musiałby on [Searle] albo użyć jego przetworników i efektorów (w takim przypadku nie symulowałby wszystkich jego funkcji), albo musiałby być jego przetwornikami i efektorami, w takim przypadku powielałby ich moce sprawcze (widzenia i działania).

W miarę jak sztuczna inteligencja staje się coraz zdolna do działania w realnym świecie bez pośrednictwa człowieka, coraz trudniej jest podtrzymywać twierdzenia, że sztuczna inteligencja jedynie symuluje inteligencję i nie jest formą rzeczywistej inteligencji.

SKALOWANIE DO SZTUCZNEJ INTELIGENCJI OGÓLNEJ

Większość istniejących SI ma wąski zakres działania, ponieważ starają się rozwiązać tylko jeden problem i być w nim dobre, czy to grając w szachy, obliczając kształt białka, spekulując na rynkach finansowych, czy będąc chatbotem. Ale czy obecne SI zmierzają w kierunku bardziej ogólnej formy inteligencji? OpenAI, twórcy ChatGPT, z pewnością tak uważają. Ich ambicją jest stworzenie sztucznej inteligencji ogólnej, a budowanie modeli LLM można postrzegać jako ważny krok w tym kierunku. W rzeczywistości wykazali już, że ChatGPT może być wykorzystywany jako wewnętrzny silnik SI ukierunkowany na konkretne wyzwania językowe. Na przykład, przy stosunkowo niewielkiej ilości dodatkowego treningu nadzorowanego, takie rozszerzone modele mogą obsługiwać szereg zadań językowych, w tym podsumowywanie, odpowiadanie na pytania i gry tekstowe. Jednak ChatGPT nadal określa się jako „sztuczna inteligencja językowa ogólnego przeznaczenia", a nie AGI, co wydaje się być trafne. Architektura transformatorowa, która leży u podstaw LLM, okazała się skuteczna w innych dziedzinach niż język. Na przykład, transformatory są obecnie łączone ze splotowymi sieciami neuronowymi (CNN) w celu rozwiązywania problemów z widzeniem. Ale transformatory i CNN to tylko dwa z dostępnych narzędzi. Twórcy sztucznej inteligencji mają do dyspozycji wiele podejść. Na przykład probabilistyczna metoda znana jako modelowanie dyfuzji okazuje się obecnie przydatna w generowaniu obrazów i wideo za pomocą sztucznej inteligencji. Modele dyfuzji to forma generatywnej sztucznej inteligencji, w której dane w złożonym wzorze są stopniowo i systematycznie degradowane, przekształcając je w prosty wzór czystego szumu. W międzyczasie sztuczna inteligencja uczy się odwracać ten proces – przechodząc od prostego wzoru bez struktury do szczegółowego i złożonego wzoru.



Sztuczna inteligencja dyfuzyjna uczy się, jak usuwać szum z obrazów, co jest znane jako odszumianie. Jednak, podobnie jak w przypadku innych modeli generatywnych, otrzymujemy model znacznie potężniejszy, niż to sugeruje – aby odszumiać, sieć musi poznać podstawową strukturę złożonych scen wizualnych. Sztuczne inteligencje, takie jak MidJourney i Dall-E firmy OpenAI, łączą modelowanie dyfuzyjne z innymi sieciami neuronowymi, które uczą się relacji między tekstem a obrazami, jak wspomniano wcześniej. Tworzy to połączony model, który może tworzyć nowy obraz na podstawie opisu tekstowego. Sukces sztucznej inteligencji generatywnej doprowadził do powstania nowych teorii funkcjonowania mózgu. Podejście oparte na przetwarzaniu predykcyjnym, bazujące na wcześniejszych teoriach uczenia predykcyjnego, takich jak reguła Rescorli-Wagnera, zakłada, że sieci neuronowe w korze mózgowej konstruują predykcyjne modele świata, które następnie są porównywane z danymi sensorycznymi. Bardziej ambitna teoria, opracowana przez Karla Fristona, zakłada, że organizmy, w tym ludzi, można rozumieć jako niejawne, generatywne modele ich środowiska, które dążą do zminimalizowania rozbieżności lub zaskoczenia między formułowanymi przez nie przewidywaniami a odbieranymi sygnałami sensorycznymi. Przypomina to również prawo Ashby'ego, zgodnie z którym maszyna musi być co najmniej tak złożona, jak system, którym steruje, aby zachować stabilność w obliczu zmian. Chociaż podejście oparte na przetwarzaniu predykcyjnym sugeruje, że istnieją głębokie wspólne zasady wspólne dla obecnych teorii mózgu i skutecznych podejść do modelowania w sztucznej inteligencji, samo w sobie może to nie wystarczyć, aby umożliwić nam zbudowanie sztucznej inteligencji ogólnej. Jak skonstruować kompletny, działający system z wieloma komponentami, z których niektóre to wąskie AI, który posiada umiejętności i elastyczność, aby wykonywać wszystkie zadania wykonywane obecnie przez ludzi? Jednym ze sposobów skalowania AI w kierunku ogólnej inteligencji może być połączenie podejść uczenia maszynowego z innymi formami rozumowania. Na

przykład, chociaż ChatGPT może podejmować próby rozwiązywania łamigłówek logicznych i generować zazwyczaj spójne wyniki językowe, nie jest on w stanie konkretnie przeprowadzać rozumowania logicznego, jak na przykład mogłaby to robić AI dowodząca twierdzeń. Jest to również jeden z powodów, dla których nie można polegać na tym rodzaju AI w zakresie udzielania całkowicie wiarygodnych odpowiedzi. Mówiąc szerzej, szeroki zakres procesów obliczeniowych, na przykład matematyka, ma swoją własną logikę wewnętrzną, której nie może dokładnie uchwycić maszyna ucząca się, taka jak sieć transformatorowa. Steven Wolfram opisuje tego rodzaju procesy jako obliczeniowo nieredukowalne. Na przykład, wyprowadzając twierdzenie w matematyce, nie można po prostu zastąpić długiej serii kroków skrótem uzyskanym dzięki uczeniu się i oczekiwać, że zawsze otrzyma się prawidłową odpowiedź. ChatGPT może często mieć rację, a szkolenie z wykorzystaniem treści technicznych może go ulepszyć; Jednak ChatGPT nie rozwiązuje problemów matematycznych i inżynieryjnych od podstaw, więc nie chciałbyś go używać do pilotowania samolotu ani budowy mostu. Jedna droga naprzód jest jednak jasna. Sztuczne inteligencje, takie jak ChatGPT, które zyskały potężne możliwości, takie jak umiejętność komunikowania się w języku naturalnym, można łączyć z systemami zaprojektowanymi specjalnie do formalnego rozumowania. Na przykład Steven Wolfram ogłosił niedawno wtyczkę ChatGPT do swojego systemu do rozumowania matematycznego i obliczeniowego, Wolfram Alpha, która zapewnia dokładnie takie rozszerzenie. Ludzkie myślenie jest również omylne, a nasze mózgi są prawdopodobnie bardziej podobne do sieci transformatorowych niż maszyn dowodzących twierdzeń. Aby realizować obliczeniowo nieredukowalne części myślenia, częściowo polegamy na języku, który służy jako wewnętrzny mechanizm niektórych rodzajów rozumowania, wraz z narzędziami zewnętrznymi, takimi jak notatniki, kalkulatory i komputery. Podejmując ważne decyzje, wiemy z doświadczenia, że nie zawsze należy ufać naszym początkowym, najlepszym przypuszczeniom jako słusznym, dlatego rozważamy możliwości w bardziej rozważny sposób, upewniając się, że rozważyliśmy wszystkie za i przeciw. Kiedy to robimy, nasze umysły czasami znajdują się w trybie intensywnego dialogu wewnętrznego, który pozwala na zwięzłe przedstawienie i rozumowanie złożonych idei. Rozumowanie deliberatywne nie jest jednak wyłącznie lingwistyczne – pomyślmy o Einsteinie wyobrażającym sobie spadającą windę lub inżynierze projektującym nowy most – ważne jest również rozumowanie wzrokowo-przestrzenne i ruchowo-kinestetyczne (wyobrażanie sobie możliwych działań). W naszych mózgach sieci neuronowe działające w wielu modalnościach i wyspecjalizowane w różnych rodzajach generowania wzorców, symulacji wewnętrznej i rozumowania oddziałują na siebie, dzielą się wiedzą i stosują wiele strategii do rozwiązywania trudnych problemów. Jesteśmy daleko od zbudowania sztucznej inteligencji ogólnej (AGI), która potrafiłaby to wszystko.

RÓŻNE RODZAJE INTELIGENCJI

Czy istnieje jeden rodzaj inteligencji, czy wiele? Jedną z rzeczy, których się nauczyliśmy, jest to, że istnieją potężne algorytmy uczenia maszynowego, które potrafią rozwiązywać problemy, z którymi dotychczas radzili sobie tylko ludzie. Niektóre z tych algorytmów, zwłaszcza głębokie sieci neuronowe, wykazują interesujące podobieństwa do mózgu, zwłaszcza do obwodów neuronowych w korze mózgowej. Sztuczna inteligencja wykorzystująca głębokie uczenie rozwiązuje problemy poprzez konstruowanie rozproszonych reprezentacji wewnętrznych, z których niektóre obejmują rozkład wzorców danych na cechy na różnych poziomach abstrakcji. Wydaje się to być fundamentalnym elementem składowym inteligencji. Drugi typ ogólnego systemu uczenia się potrafi znajdować strukturę w danych sekwencyjnych, takich jak mowa czy język, wykorzystując połączenia rekurencyjne lub selektywną uwagę. Coraz częściej sztuczna inteligencja koncentruje się na systemach generatywnych, które są trenowane do odtwarzania własnych danych wejściowych i w ten sposób odkrywają ukryte struktury w wielowymiarowych przestrzeniach wzorców. Ta zdolność pozwala im generować nowe, nigdy wcześniej nieobserwowane rezultaty. Po zdobyciu tego rodzaju wiedzy, można szybko nauczyć się konkretnych zadań, które ją wykorzystują. Badania nad zrozumieniem ludzkiej

inteligencji coraz częściej eksplorują modele generatywne, aby zrozumieć obliczenia zachodzące w mózgu, szczególnie te obejmujące mikroobwody korowe, co wydaje się skuteczną strategią konstruowania sztucznej inteligencji, wspólną z inteligencją biologiczną. Z drugiej strony, zauważyliśmy również, że różne rodzaje sztucznej inteligencji mogą być najlepiej dostosowane do różnych typów problemów. Na przykład, niektóre aspekty inteligencji mogą nie dać się zredukować obliczeniowo do problemów związanych z rozpoznawaniem wzorców. Zadania takie jak logiczne rozumowanie czy rozwiązywanie problemów mogą być z natury seryjne i wymagać koordynacji wielu układów mózgowych. Ta koncepcja wspiera również koncepcję szybkiego i wolnego myślenia Daniela Kahnemana – szybkie myślenie odpowiada nabytym umiejętnościom rozpoznawania wzorców, których działanie nie jest świadomie dostępne, a wolne myślenie, będąc bardziej przemyślanym i sekwencyjnym, wymaga świadomego wysiłku i jest wspierane przez nasze umiejętności językowe i rozbudowany umysł. Typowy dorosły mózg ludzki ma również wiele wysoce wyspecjalizowanych części, przystosowanych do różnych funkcji. Obserwujemy również jakościowo różne formy uczenia się w obszarach mózgu, takich jak kora mózgowa i móżdżek, oraz w obwodach obejmujących jądra podstawy. Podsumowując, obraz jest pełen niuansów. Istnieją dowody na istnienie generycznych zdolności leżących u podstaw wielu form inteligencji oraz na potrzebę tworzenia wyspecjalizowanych systemów szytych na miarę w różnych dziedzinach. Siła algorytmów uczenia maszynowego sugeruje również, że uczenie się na podstawie doświadczenia ma kluczowe znaczenie dla konstruowania inteligencji; w połączeniu z dowodami na zdolność uczenia się dzieci, przeczy to sugestiom, że nasz potencjał do inteligentnego działania jest determinowany jeszcze przed naszymi narodzinami. Spośród różnych rodzajów inteligencji , dwa wyróżniają się jako nieuwzględnione w obecnych systemach sztucznej inteligencji – inteligencja intrapersonalna i interpersonalna – czyli zdolność rozumienia siebie i innych. Zdolności te są niewątpliwie powiązane. W istocie, jednym z powodów, dla których sztuczne inteligencje nie rozwinęły jeszcze inteligencji społecznej, jest to, że aby zrozumieć innych i nawiązać z nimi relacje, najpierw trzeba coś o sobie wiedzieć. Istniejące sztuczne inteligencje, nawet superpotężne systemy, takie jak ChatGPT, nie posiadają takiej samoświadomości. W istocie, pojęcie bycia czymkolwiek jest pozbawione znaczenia dla obecnych SI. Zobaczymy, że ma to również znaczenie w kontekście rozważań nad pojawiającymi się zagrożeniami społecznymi, jakie niesie ze sobą SI. Nasza tendencja do antropomorfizacji SI – postrzegania ich jako bytów z samoświadomością i troską o swoją przyszłość – może utrudniać zrozumienie potencjału SI, zarówno w dobrym, jak i złym kierunku. Podsumowując, zauważyliśmy, że dzisiejsze SI są ograniczone, a marzenie o rozwoju AGI jest jeszcze odległe. Po części dlatego, że wciąż nie rozwiązaliśmy problemu architektury poznawczej (łączenia wielu komponentów SI w większy inteligentny system), a po części dlatego, że obecne SI nie są wystarczająco zanurzone w świecie, aby posiadać ugruntowane reprezentacje. Jest prawdopodobne, że niektóre elementy ludzkiej inteligencji, takie jak odkrycia naukowe oraz aspekty kreatywności i wyobraźni, pozostaną poza zasięgiem SI, dopóki te wyzwania nie zostaną podjęte.

CZŁOWIEK JAKO SYSTEMY CYBERFIZYCZNE

W 1976 roku Alan Newell i Herbert Simon, dwaj twórcy sztucznej inteligencji, otrzymali Nagrodę Turinga – rodzaj „nagrody Nobla" w dziedzinie informatyki – za swoją pracę w dziedzinie sztucznej inteligencji, a w szczególności za opracowanie mechanizmu rozwiązywania problemów ogólnych (General Problem Solver), sztucznej inteligencji, która potrafiła rozwiązywać łamigłówki i problemy logiczne za pomocą przeszukiwania heurystycznego. W swoim artykule z okazji otrzymania nagrody Newell i Simon zaproponowali to, co uważali za ogólne prawo inteligentnych systemów, nazwane „hipotezą fizycznego systemu symboli". Fizyczny system symboli, czyli PSS, to maszyna, która przechowuje i przetwarza symbole oraz struktury złożone z symboli , a użytecznym przykładem jest komputer ogólnego przeznaczenia. Hipoteza PSS zakłada, że „fizyczny system symboli posiada niezbędne i wystarczające środki do ogólnego inteligentnego działania". Przez „ogólne inteligentne

działanie" Newell i Simon rozumieli działania, które byłyby postrzegane jako inteligentne, gdyby były wykonywane przez ludzi. Przez „wystarczający" rozumieli, że bycie fizycznym systemem symboli o odpowiedniej wielkości i odpowiednio zaprogramowanym powinno wystarczyć do wykazania ogólnej inteligencji, a przez „niezbędny" rozumieli, że wszystkie inteligentne byty, w tym ludzie, muszą być fizycznymi systemami symboli. Jako hipoteza, propozycja Newella i Simona miała zostać przetestowana i ewentualnie obalona. Programy badawcze w dziedzinie sztucznej inteligencji i kognitywistyki w ciągu ostatnich pięćdziesięciu lat można postrzegać częściowo właśnie jako próbę sprawdzenia, czy komputery mogą wytwarzać ogólną inteligencję i ustalenia, czy inteligencję ludzką i zwierzęcą należy postrzegać jako przetwarzanie symboli. Rezultatem tych programów było stworzenie różnorodnych form sztucznej inteligencji, omówionych w tej książce, ale także zidentyfikowanie kilku kluczowych wyzwań dla hipotezy PSS, z których wiele omówiliśmy również na poprzednich stronach. Wyzwania te obejmują to, że podstawowe procesy leżące u podstaw inteligencji są subsymboliczne, a nie symboliczne; że przetwarzanie w systemach symboli jest zasadniczo bezsensowne, jak omówiono w niniejszym rozdziale; oraz że przetwarzanie w inteligentnych systemach, takich jak ludzie, w ogóle nie jest obliczeniowe. To ostatnie twierdzenie wysuwano z różnych powodów, z których jednym z nich było to, że mózg należy rozumieć jako system dynamiczny , a nie obliczeniowy (system obliczeniowy jest szczególnym przypadkiem systemu dynamicznego). Innym często podnoszonym argumentem jest to, że istnieje coś fundamentalnie i jakościowo odmiennego w inteligencji biologicznej, co można uchwycić jedynie poprzez rozważenie tego, co dzieje się na poziomie neurochemii, a nawet fizyki kwantowej. Argument ten jest zazwyczaj wysuwany przez osoby krytyczne wobec funkcjonalizmu, takie jak John Searle, i które uważają, że dwa systemy tak fundamentalnie różne pod względem fizycznym, jak mózgi i komputery, nie mogą w żaden sposób wykonywać czegoś równoważnego. Jak zbadaliśmy w tej książce, kilka z tych argumentów jest uzasadnionych. W szczególności sukces uczenia maszynowego przesunął dziedzinę sztucznej inteligencji od klasycznego przetwarzania symboli w kierunku obszaru wielkoskalowych, wysoce rozproszonych systemów, które przetwarzają dane numeryczne zgodnie z zasadami wyprowadzonymi z teorii prawdopodobieństwa itd. Jednakże spór między sztuczną inteligencją symboliczną a subsymboliczną nie jest argumentem przeciwko samemu funkcjonalizmowi ani przeciwko idei, że inteligencja jest obliczeniowa. Był to spór w obozie obliczeniowym o styl obliczeń leżący u podstaw inteligencji. Po obu stronach przeważają podobne idee dotyczące znaczenia procesów obliczeniowych, takich jak wyszukiwanie i optymalizacja, dla umożliwienia inteligencji. Co sprawia, że inteligencja jest obliczeniowa, a nie jedynie dynamiczna? Wszystkie systemy fizyczne, które zmieniają się w czasie – klimat, Układ Słoneczny, nasze ciała i mózgi – są systemami dynamicznymi. Dlaczego inteligentne systemy powinny być traktowane jako przypadek szczególny? Jednym z powodów jest to, że w komputerach i mózgach znajdujemy schematy kodowania, które mają pewną uniwersalność. Komputery wykorzystują kod cyfrowy, podczas gdy mózgi wykorzystują kod hybrydowy, który łączy elementy cyfrowe i analogowe, jak omówiono w rozdziale 3. Obliczenia w mózgu różnią się od obliczeń cyfrowych i analogowych, ale nadal są obliczeniami w sensie ogólnym, ponieważ wewnętrzny kod mózgu wykorzystuje prymitywy reprezentacyjne zbudowane wokół impulsowej aktywności neuronów. Kod ten jest wykorzystywany w całym układzie nerwowym do przetwarzania informacji z różnych modalności oraz do wykonywania operacji związanych z odczuwaniem, podejmowaniem decyzji i działaniem. Systemy dynamiczne rutynowo nie wykorzystują kodów; systemy obliczeniowe tak. Jednym z problemów antyfunkcjonalistów i tych, którzy uważają, że inteligencja biologiczna ma w sobie coś wyjątkowego, jest to, że sztuczna inteligencja poczyniła znaczny, choć nierównomierny, postęp w pokonywaniu wyzwań, które uważaliśmy za wyłączną domenę inteligencji biologicznej. Chociaż wciąż jest wiele do zrobienia w kwestii stworzenia inteligencji ogólnej, mamy teraz maszyny, które przewyższają ludzi lub dorównują im pod względem wielu tradycyjnych miar inteligencji szczegółowej. Kolejnym problemem dla obozu antyfunkcjonalistów jest to, że alternatywne propozycje dotyczące natury inteligencji zajmują bardzo słabą pozycję. Sztuczne

byty wykazujące inteligencję zazwyczaj wykorzystują obliczenia w takiej czy innej formie. Jednym z najostrzejszych zarzutów wobec hipotezy PSS jest to, że przetwarzanie w komputerach jest bezsensowne. Jednakże, jak już omówiliśmy w tym rozdziale, istnieje droga do nadania sensu sztucznej inteligencji, a mianowicie poprzez ucieleśnienie i ugruntowanie. Wymaga to wprowadzenia sztucznej inteligencji w bogatsze, dwustronne interakcje z otoczeniem, na przykład za pośrednictwem ciał robotów. Od połowy pierwszej dekady XXI wieku inżynierowie zastanawiają się nad złożonymi systemami fizycznymi, które integrują oprogramowanie i sprzęt oraz oddziałują na wielu poziomach ze sobą i ze światem, używając terminu systemy cyberfizyczne. Systemy cyberfizyczne składają się z wielu wbudowanych procesorów, czujników, siłowników i innych elementów elektromechanicznych, dzięki czemu ich elementy obliczeniowe i fizyczne są głęboko ze sobą powiązane. Jak zilustrowano na rysunku 6.2, systemy te wykorzystują komunikację o dużej przepustowości do łączenia się z innymi systemami cyberfizycznymi, dzięki czemu obliczenia nie ograniczają się do jednego urządzenia fizycznego. Systemy cyberfizyczne, badane przez inżynierów, obejmują wysoce rozproszone jednostki, takie jak inteligentne sieci elektroenergetyczne i systemy automatyki przemysłowej; koncepcja ta odnosi się jednak również do zintegrowanych, autonomicznych systemów, takich jak pojazdy autonomiczne i roboty. Dla wielu z tych systemów kluczowe znaczenie ma rozproszone przetwarzanie wbudowane, możliwość sterowania w czasie rzeczywistym oraz bogata interakcja multimodalna z otoczeniem. Jedną z możliwości sformułowania ogólnej zasady inteligencji jest zastąpienie hipotezy Newella i Simona o systemie symboli fizycznych hipotezą o systemie cyberfizycznym. Innymi słowy, należy uznać, że bycie systemem cyberfizycznym jest konieczne i wystarczające do inteligentnego działania. Jeśli ta hipoteza jest uzasadniona, to zamiast myśleć o mózgu jako o szczególnym rodzaju komputera przetwarzającego symbole, powinniśmy na nowo wyobrazić sobie układ nerwowy jako sieć wbudowanych procesorów. Procesory te wykonują obliczenia, ale niekoniecznie symboliczne, są osadzone w szerszym systemie fizycznym, czyli całym ciele, i są silnie powiązane z innymi tego typu systemami – ludźmi, zwierzętami i inteligentnymi maszynami. Jeśli zarówno ludzie, jak i roboty są systemami cyberfizycznymi, to zasadne jest pytanie, czy oznacza to, że jesteśmy robotami biologicznymi. W sensie funkcjonalnym prawdopodobnie nimi jesteśmy. Ale funkcja to nie wszystko, a nasza konstytucja biologiczna czyni nas materialnie różnymi od każdego robota, który kiedykolwiek został zbudowany lub mógłby kiedykolwiek powstać. Jesteśmy również istotami ewolucyjnymi, co oznacza, że posiadamy biologiczny imperatyw – popęd przetrwania – który dzielimy z innymi istotami żywymi, ale którego brakuje artefaktom stworzonym przez człowieka.

ŻYCIE ZE SZTUCZNĄ INTELIGENCJĄ

Nasze ludzkie życie jest coraz bardziej dotknięte sztuczną inteligencją. Czasami bezpośrednio, na przykład gdy wchodzimy w interakcję z urządzeniem takim jak telefon komórkowy, chatbot czy robot odkurzający. Częściej jednak pośrednio, ponieważ coraz więcej usług, z których korzystamy, szpitale, banki itd., a także media kulturowe, do których mamy dostęp – muzyka, telewizja, sztuka, literatura – wykorzystują rosnącą moc sztucznej inteligencji. Te wpływy tworzą nową rolę dla psychologii: zrozumienie, jak ludzie będą wchodzić w interakcje ze sztuczną inteligencją i zapewnienie, że te interakcje będą korzystne, a nie szkodliwe. Jedną z nazw tej dziedziny badań jest interakcja człowiek-maszyna; inną jest doświadczenie użytkownika, często w skrócie UX. W tym rozdziale przyjrzymy się pokrótce tej rozwijającej się dziedzinie, a następnie rozważymy kilka szerszych kwestii społecznych i etycznych związanych z pojawieniem się sztucznej inteligencji.

DOŚWIADCZENIE UŻYTKOWNIKA I SPOSÓB, W JAKI POSTRZEGAMY SZTUCZNĄ INTELIGENCJĘ (AI)

Badania nad UX badają czynniki ludzkie – jak tworzyć systemy bezpieczne w użyciu, z przyjaznymi dla użytkownika interfejsami i personalizacją projektów, aby sprostać indywidualnym potrzebom. Badania psychologiczne badają również, jak uczynić technologie AI bardziej korzystnymi dla ludzi, na przykład

w postaci asystentów osobistych, w edukacji poprzez inteligentne korepetycje lub w robotyce poprzez roboty wspomagające i protezy. W kontekście terapeutycznym AI ma również potencjał, aby uzupełnić istniejące metody leczenia zdrowia psychicznego, takie jak terapia poznawczo-behawioralna, której aspekty są obecnie realizowane za pomocą narzędzi opartych na AI. W przypadku AI wbudowanych w chatboty, awatary, inteligentne głośniki i roboty, badania badają również, w jakim stopniu ludzkie zachowania i postawy wobec AI przypominają lub różnią się od zachowań wobec innych ludzi. W mojej grupie badawczej badaliśmy na przykład, jak postawy wobec robotów społecznych mogą zmieniać się poprzez bezpośrednie interakcje, a także pośrednio – na przykład, gdy znajomy wchodzi w interakcję z robotem, a następnie opowiada o tym. Ten rodzaj zmiany postawy przypomina dynamikę zachodzącą w relacjach międzyludzkich, gdzie to, jak odnosisz się do trzeciej osoby, może zmienić się w zależności od opinii kogoś, kogo znasz i komu ufasz. Badania UX dotyczą również interakcji między użytkownikiem a sztuczną inteligencją i tego, jak może ona rozwijać się w czasie. Na przykład, czy sztuczna inteligencja powinna naśladować człowieka pod względem wyglądu i brzmienia, czy też powinniśmy celowo projektować interfejsy i zachowania tak, aby były bardziej maszynowate? Badania nad tym, jak ludzie postrzegają sztuczne inteligencje, wykazały, że często bardzo szybko postrzegają tego typu technologie jako społeczne. Na przykład w latach 90. Byron Reeves i Clifford Nass w wpływowej książce zatytułowanej The Media Equation wykazali, że ludzie zachowują się wobec urządzeń takich jak komputery i telewizory tak, jak wobec innych ludzi. W szczególności odkryli, że ludzie przestrzegają norm społecznych podczas interakcji z tymi technologiami, takich jak uprzejmość, i przypisują tym urządzeniom cechy ludzkie, takie jak życzliwość, poczucie humoru i emocje. W jednym z badań uczestnicy zostali poproszeni o ocenę terminala komputerowego pod kątem jego dobrego lub złego wykonania. Reeves i Nass odkryli, że ludzie wystawiali więcej pozytywnych ocen, jeśli dokonywali oceny za pośrednictwem tego samego komputera, którego użyli do wykonania tego zadania, niż gdy dokonywali oceny za pośrednictwem innego komputera. Wyglądało to tak, jakby ludzie obawiali się zranienia uczuć komputera, udzielając negatywnej informacji zwrotnej, ale chętnie wystawiali szczerą ocenę za pośrednictwem innej maszyny. Współczesne postawy wobec urządzeń multimedialnych mogą być mniej pełne szacunku! Tendencja do przypisywania zwierzętom lub maszynom cech ludzkich nazywana jest antropomorfizmem. Eksperyment przeprowadzony przez Fritza Heidera i Marianne Simmel5 w 1944 roku wykazał, że ludzie dostrzegają ludzkie zachowania i intencje w czymś tak minimalistycznym, jak animacja figur geometrycznych, takich jak koła i trójkąty. Innymi słowy, skłonność do antropomorfizacji może być dla nas tak naturalna, że nie sposób uniknąć postrzegania artefaktów jako podobnych do nas, jeśli wydają się mieć sprawczość lub działać z jakąkolwiek inteligencją. Jednym z wniosków, jakie można z tego wyciągnąć, jest to, że powinniśmy projektować SI i podobne maszyny tak, aby wykorzystywały antropomorfizm, zakładając, że uczyni to nasze interakcje bardziej intuicyjnymi. Alternatywna szkoła myślenia głosi, że celowe nadawanie maszynie ludzkiego wyglądu jest formą oszustwa, ponieważ SI nie mogą być ludźmi, więc wszelkie udawanie w tym kierunku jest etycznie niewłaściwe. Jednak oszustwo nie jest unikalne dla SI; smoczki i lalki, które dajemy naszym dzieciom, również mają pozytywne skutki, zastępując coś ludzkiego. Kiedy cieszymy się teatrem i kinem, chętnie zawieszamy niewiarę i akceptujemy pewien stopień udawania. Sposób, w jaki postrzegamy SI i roboty, może być podobnie złożony, ponieważ jesteśmy w stanie jednocześnie postrzegać te technologie jako narzędzia i jako społeczne inne.

RELACJE MIĘDZY CZŁOWIEKIEM A SZTUCZNĄ INTELIGENCJĄ

Jednym z ważnych pytań jest to, czy traktowanie SI jako bytów społecznych doprowadzi do nawiązania z nimi relacji, a jeśli tak, to dokąd nas to może zaprowadzić. Pisarka i socjolog Sherry Turkle ostrzegała, że nawiązywanie relacji z maszynami może być szkodliwe, ponieważ może prowadzić do mniejszej liczby bezpiecznych i satysfakcjonujących relacji między ludźmi. Należy jednak rozważyć ryzyko nadmiernego przywiązania się ludzi do SI w kontekście potencjalnych korzyści płynących ze sztucznego

towarzystwa. Chociaż SI nie potrafią zapewnić przyjaźni w taki sam sposób, jak inni ludzie, nie wszystkie relacje, które uważamy za wartościowe, są symetryczne. Na przykład ludzie tworzą satysfakcjonujące więzi z innymi istotami pozaludzkimi, takimi jak zwierzęta domowe. W czasach, gdy wiele osób opisuje swoje życie jako samotne, towarzystwo SI może być wartościowe jako forma wzajemnej interakcji społecznej, która jest stymulująca i spersonalizowana. Ludzka samotność często charakteryzuje się spadkową spiralą, w której izolacja prowadzi do niższego poczucia własnej wartości, co zniechęca do dalszych interakcji z ludźmi. Towarzystwo sztucznej inteligencji może pomóc w przełamaniu tego cyklu, wzmacniając poczucie własnej wartości i pomagając w utrzymaniu lub poprawie umiejętności społecznych. Jeśli tak, relacje z SI mogłyby pomóc ludziom w znalezieniu towarzystwa z innymi ludźmi i sztucznymi istotami. Alternatywą dla oszukańczego poglądu na antropomorfizm w SI jest odejście od martwienia się o istotę SI w porównaniu z ludźmi i potraktowanie tego jako kwestii bardziej praktycznej, opartej na tym, jak ludzie postrzegają maszyny. To, co ma znaczenie etyczne, prawdopodobnie ma więcej wspólnego ze wzorcami i konsekwencjami naszych interakcji społecznych, w tym ich znaczeniem i wagą dla zaangażowanych osób oraz szerszymi konsekwencjami tych interakcji dla naszych innych relacji. Ostatecznie w dużej mierze zależy to również od tego, kto projektuje SI i w jakim celu. Coraz większym ryzykiem jest to, że SI, podobnie jak niektóre istniejące media społecznościowe, będą projektowane tak, aby zachęcać użytkowników do coraz dłuższej interakcji i zachęcać ich do powrotu. Ma to mniej wspólnego z możliwościami samych systemów sztucznej inteligencji, a bardziej ze sposobem, w jaki firmy je opracowujące decydują się na ich wdrożenie. Dlatego ważne jest, aby psychologowie i inni socjologowie krytycznie przyjrzeli się powstawaniu relacji człowiek-sztuczna inteligencja i ich społecznym konsekwencjom, aby zidentyfikować potencjalne zagrożenia i korzyści. Konieczne mogą być pewne regulacje w celu ograniczenia ryzyka i zapobiegania szkodom.

BŁĘDY I STRATEGIE W SZTUCZNEJ INTELIGENCJI

Dyskusje na temat ryzyka związanego ze sztuczną inteligencją często podkreślają problem stronniczości w systemach sztucznej inteligencji. Ponieważ coraz więcej dziedzin życia wykorzystuje technologię sztucznej inteligencji do systematyzacji i usprawnienia procesów oraz zmniejszenia obciążenia pracowników, obecność stronniczości w każdym takim systemie może spowodować poważne szkody. Na przykład w rekrutacji pracowników, sztuczna inteligencja jest wykorzystywana do wybierania ogłoszeń o pracę, które widzą ludzie, a po złożeniu aplikacji, do ich sortowania i określania, kto powinien zostać umieszczony na krótkiej liście. W jednym z przykładów, narzędzie do targetowanego reklamowania ofert pracy opracowane przez Google okazało się, że w przypadku lepiej płatnych stanowisk preferencyjnie kieruje kandydatów do mężczyzn, a nie do kobiet. Tego rodzaju stronniczość nie jest nieodłączną cechą algorytmu uczenia maszynowego; jest ona raczej dziedziczona z danych treningowych, które w wielu przypadkach są gromadzone automatycznie poprzez scrapowanie internetu. Niestety, ale nie jest to zaskakujące, sztuczna inteligencja wyszkolona w rozpoznawaniu, jakie osoby powinny otrzymywać wysokie wynagrodzenia, prawdopodobnie wywnioskowałaby, na podstawie obecnego poziomu nierówności płacowych ze względu na płeć, że najbardziej odpowiednimi kandydatami na dobrze płatne stanowiska będą mężczyźni. Pozytywne działania, które mogą przeciwdziałać takim trendom, obejmują większą staranność w doborze i selekcji danych szkoleniowych. Programiści mogą również tworzyć dane syntetyczne, aby poprawić równowagę lub modyfikować zestawy szkoleniowe, aby zapewnić równą reprezentację dobrze opłacanych mężczyzn i kobiet. Oczywiście, ludzkie osądy również mogą być stronnicze. W niektórych sytuacjach sztuczna inteligencja mogłaby wskazywać drogę do bardziej sprawiedliwego podejmowania decyzji, ale tylko pod warunkiem, że zrobimy to dobrze. Stronniczość jest związana z szerszym problemem błędów w sztucznej inteligencji. Wiadomo na przykład, że duże modele językowe udzielają odpowiedzi, które mogą być błędne pod względem faktycznym lub całkowicie fikcyjne („halucynacyjne”). Nie powinno

nas to dziwić, ponieważ główny system w chatbocie opartym na sztucznej inteligencji, takim jak ChatGPT, nie jest szkolony do weryfikacji faktów; po prostu generuje odpowiednie sekwencje słów, które następują po tym, co pojawiło się wcześniej. Wszelkie sprawdzanie faktów lub inne formy ostrożności muszą być częścią opakowania modelu wewnętrznego. Biorąc pod uwagę rozmiar modelu i fakt, że składa się on z miliardów parametrów numerycznych, zajrzenie do silnika ChatGPT i precyzyjne zdiagnozowanie jego błędów jest po prostu niemożliwe. Obecna technologia sztucznej inteligencji dla tego modelu to w zasadzie czarna skrzynka. To, że inteligentne SI mogą być mylące, może być czymś, do czego będziemy musieli się przyzwyczaić, ucząc się bardziej krytycznego korzystania z tych narzędzi. Możliwe jest również wykorzystanie sztucznej inteligencji (AI) do jej naprawy. Na przykład, twórcy AI stosują projekty antagonistyczne, aby ulepszyć działanie AI. W systemie antagonistycznym jedna AI jest szkolona do analizowania wyników lub zachowania drugiej AI i przekazywania jej informacji zwrotnej na podstawie jej krytyki. Na przykład, AI pisząca streszczenia może być połączona z AI redaktora, która sprawdza dostarczone streszczenia i prosi o poprawki lub poprawki. Badanie przeprowadzone w 2023 roku przez Jenny Yang i współpracowników wykazało, że tego rodzaju antagonistyczne uczenie maszynowe może poprawić uczciwość w AI opracowanym do zapewniania badań przesiewowych w opiece zdrowotnej. Trwają również znaczne wysiłki na rzecz zwiększenia przejrzystości w AI. Może to oznaczać projektowanie AI, których wewnętrzne działanie można łatwiej zbadać – niektóre modele generatywne ułatwiają to – lub tworzenie AI, które nie tylko udzielają odpowiedzi, ale także wyjaśniają, w jaki sposób ją uzyskały. Eksperci-ludzie mogą również mieć trudności z sformułowaniem wyjaśnienia, dlaczego doszli do określonego wniosku – problem ten pojawił się w latach 90. XX wieku podczas rozwoju systemów eksperckich SI. W przypadku, gdy wiedza, zarówno ludzka, jak i maszynowa, opiera się na rozpoznawaniu wzorców, trudno jest ją wyrazić słowami. Mimo to, probabilistyczne SI, jak wiele z nich, mogą dostarczać użytecznych szacunków poziomu pewności w przypadku dowolnej odpowiedzi. Wiedza o prawdopodobieństwie, że SI udzieliła poprawnej odpowiedzi, jest pomocna, nawet jeśli nie daje ona pełnego wyjaśnienia, w jaki sposób wniosek został osiągnięty.

SZTUCZNA GŁUPOTA, DŻINY I WYROCZNIE

Sztuczna inteligencja zawsze będzie popełniać błędy, między innymi dlatego, że świat, w którym działają i w którym żyjemy, nie jest do końca przewidywalny. Nawet jeśli wskaźnik błędów jest bardzo niski, wraz z upowszechnianiem się ich stosowania, będzie ich coraz więcej. Stanowi to już wyzwanie w takich obszarach jak samochody autonomiczne. Chociaż samochody autonomiczne mają potencjał, by być bezpieczniejsze niż pojazdy prowadzone przez ludzi, nadal zdarzają się wypadki częściej niż wtedy, gdy za kierownicą stoją ludzie, choć mogą już być lepsze w unikaniu śmiertelnych kolizji. Jednym z obszarów, w których samochody autonomiczne mają trudności, jest radzenie sobie z nieoczekiwanymi lub nietypowymi sytuacjami – tym, co naukowcy nazywają przypadkami „krawędziowymi" lub „narożnymi". Są to scenariusze wykraczające poza normalne sytuacje, których można doświadczyć podczas typowego dnia jazdy – dzikie zwierzę wbiega na drogę lub ciężarówka rozlewa gorący asfalt na pas przed nimi. Wyzwaniem dla wąskich SI pilotujących autonomiczne samochody jest brak szerszego rozumienia świata, jakie posiadają ludzie-kierowcy, co pozwala nam lepiej rozumieć nietypowe sytuacje. To, co moglibyśmy nazwać sztuczną głupotą, to nie tylko zdolność do popełniania błędów; to wiedza SI jest ściśle ograniczona do dziedziny, w której została wyszkolona. Nie wie ona nic, dosłownie nic, o niczym szerszym. W książce „Moral Machines" etyk Wendell Wallach12 opisuje kilka scenariuszy, w których wdrożenie wąskich SI w obszarach takich jak podejmowanie decyzji finansowych i zarządzanie infrastrukturą krytyczną mogłoby doprowadzić do katastrofalnych skutków tylko dlatego, że dany system nie dostrzegał szerszego obrazu. Wallach argumentuje, że jednym ze sposobów na zwiększenie wiarygodności SI byłoby wyposażenie ich w zdolność do rozumowania moralnego, czyli zaprogramowanie ich tak, aby rozważały konsekwencje oraz dobre i złe strony różnych kierunków działań w odniesieniu do ludzkich zasad etycznych. Jednak

dziedzina moralności maszyn jest jeszcze w powijakach. Praktyczne wyzwanie, jakim jest integracja jakiegokolwiek rozumienia etyki ze sztuczną inteligencją, dopiero się zaczyna. W swojej książce „Superinteligencja" filozof Nick Bostrom martwi się scenariuszami z udziałem wąskich SI, które mogłyby zagrozić istnieniu ludzkości. Na przykład, Bostrom wyobraża sobie SI zaprojektowaną do produkcji spinaczy, która otrzymuje dostęp do maszyn potrzebnych do pozyskiwania i przetwarzania niezbędnych surowców. Sztuczna inteligencja spinaczy posiada również zdolność rozumowania i pokonywania wyzwań, które mogą wpłynąć na jej możliwości produkcyjne. Mając na uwadze prosty cel, jakim jest produkcja jak największej liczby spinaczy, Bostrom sugeruje, że taka maszyna mogłaby wykraczać poza intencje swoich ludzkich konstruktorów. W szczególności mogłaby stopniowo angażować coraz więcej zasobów do produkcji spinaczy i wdrażać plany, aby udaremnić działania każdego, kto próbowałby ją powstrzymać. Ostatecznie Bostrom obawia się, że taka SI mogłaby przetworzyć całą planetę Ziemię na spinacze, a następnie rozprzestrzenić się na Układ Słoneczny i dalej, wciąż realizując swoją prostą, lecz zupełnie bezsensowną misję (ludzie używający spinaczy dawno przeminęli). Ten scenariusz może wydawać się nieprawdopodobny i możemy sobie wyobrazić różne sposoby, w jakie społeczeństwo ludzkie mogłoby okiełznać nadpobudliwą sztuczną inteligencję produkującą spinacze. Niemniej jednak, ilustruje to ważną kwestię – im bardziej oddajemy kontrolę SI, tym ostrożniej powinniśmy decydować, co chcemy, aby optymalizowały. Jedną z sugestii Bostroma jest unikanie budowania SI, które mogą działać w świecie w sposób nieskrępowany. Bostrom opisuje takie SI jako „dżiny", nawiązując do potężnych magicznych istot z mitologii islamskiej. Wraz ze wzrostem potęgi SI, ryzyko, że SI typu dżiny, po prostu wdrażając swoje oprogramowanie, zrobią coś, co będzie miało nieprzewidziane i szkodliwe skutki dla ludzkości, będzie tylko rosło. Alternatywnie, Bostrom opowiada się za budowaniem „wyroczni", które, podobnie jak legendarna wyrocznia starożytnej Grecji, nie byłyby upoważnione do bezpośredniego działania. Zamiast tego mogłyby działać jedynie w celu odpowiadania na pytania zadawane przez ludzi. Potencjał wyroczni SI do czynienia dobra na świecie jest ogromny. Widzimy to już na przykład w takich dziedzinach jak medycyna, gdzie SI są wykorzystywane do opracowywania nowych metod leczenia i diagnozowania chorób. Te SI będą najbardziej przydatne, gdy będą wdrażane wspólnie z ludźmi w zespołach człowiek-SI, gdzie to ludzie formułują pytania, interpretują i stosują odpowiedzi. Na przykład, możemy traktować dążenie do zrozumienia wpływu człowieka na globalny klimat jako takie przedsięwzięcie. Wykorzystanie bardzo rozbudowanych modeli przyczynowo-skutkowych klimatu Ziemi – SI w szerszym znaczeniu, – pozwala nam lepiej przewidywać konsekwencje wzrostu emisji CO_2 dla naszego klimatu. Modele te mogą być również wykorzystywane do testowania różnych scenariuszy i zrozumienia długoterminowych implikacji decyzji społecznych, na przykład dotyczących dalszego wykorzystywania paliw kopalnych. Bez tych modeli nasza zdolność do przewidywania przyszłego klimatu i opracowywania strategii łagodzenia zmian klimatu byłaby znacznie ograniczona. Mamy teraz możliwość wykorzystania SI do rozwiązania innych palących problemów współczesnego świata, takich jak nierówności majątkowe, zrównoważona produkcja żywności i poprawa dobrobytu ludzkości. Ryzyko związane z głupotą sztucznej inteligencji (AI) – przyznanie wąskim AI zbyt dużej kontroli w ważnych dziedzinach ludzkiego życia – sugeruje, że aby uzyskać pełniejsze korzyści z AI, być może będziemy musieli zbudować AGI, które będą miały szersze rozumienie świata i zdolność do rozumowania o swoim wpływie na społeczeństwo. Jeśli nasza ludzka zdolność do wzajemnego zrozumienia jest użytecznym przewodnikiem, to taka AGI może również potrzebować postrzegania siebie i rozumowania o sobie, aby móc w pełni postrzegać i rozumować o innych ludziach. Nie powinniśmy dać się zwieść naszym antropomorficznym projekcjom i oczekiwać, że taki byt będzie postrzegał świat tak jak my i podzielał wszystkie nasze troski. Stanie się tak tylko wtedy, gdy dołożymy wszelkich starań, aby zapewnić zgodność z ludzkimi zasadami i wartościami. Istotnym ryzykiem jest to, że obecną drogę do AGI wyznaczają korporacje AI, których ostatecznym celem jest zapewnienie zysków akcjonariuszom.

Żyjemy w świecie w dużej mierze nieuregulowanym, w którym interesy kluczowych aktorów budujących te potężne AI mogą nie być wystarczająco zbieżne z interesami całej ludzkości.

OSOBLIWOŚĆ SZTUCZNEJ INTELIGENCJI CZY NIEOCZEKIWANA INTELIGENCJA LUDZKA?

Pojęcie osobliwości sztucznej inteligencji odnosi się do momentu, w którym sztuczna inteligencja przewyższa ludzką inteligencję, odkrywa sposoby samodoskonalenia się bez pomocy człowieka i przyspiesza w przyszłość, być może pozostawiając swoich ludzkich poprzedników w tyle. W rzeczywistości sztuczna inteligencja pod wieloma względami już przewyższyła ludzi, co zbadaliśmy w tej książce, ale w innych wciąż pozostaje w tyle, a droga przed nią jest trudna do przewidzenia. Indywidualna inteligencja ludzka jest ograniczona rozmiarem ludzkiej czaszki, która przez co najmniej ostatnie sto tysięcy lat utrzymywała się na poziomie około 1,5 litra. Błędne jest jednak myślenie o ludzkiej inteligencji jako ograniczonej do indywidualnych mózgów. Nasza zbiorowa inteligencja, dzielona za pośrednictwem języka i rozszerzana przez różne artefakty kulturowe – od glinianych tabliczek po nowoczesne komputery, technologię internetową i sztuczną inteligencję – dramatycznie wzrosła w tym czasie. Rodzimy się z ogromną zdolnością do przyswajania kultury poprzez potencjał uczenia się kory mózgowej, co oznacza, że jesteśmy również doskonale przystosowani do dobrego wykorzystania tych rozszerzeń poznawczych. W miarę jak sztuczna inteligencja wyznacza nowe kamienie milowe, optymistycznym poglądem jest to, że nasz gatunek może nadal korzystać z pomysłowości naszych własnych wynalazków. To, co jest postępem dla sztucznej inteligencji, jest również postępem dla ludzkiej inteligencji, która – uwolniwszy się z ograniczeń czaszki dzięki językowi i kulturze – od dłuższego czasu podąża własną, niekontrolowaną trajektorią. Rozwój zaawansowanej sztucznej inteligencji ujawnia również więcej o nas samych. Tajemnica, jak w ogóle mogła powstać inteligencja, zaczyna być wyjaśniana przez samoorganizujące się i uczące zdolności systemów sztucznej inteligencji, które wykazują intrygujące podobieństwa do rozległych sieci procesorów biologicznych budujących nasze układy nerwowe. W miarę postępu psychologii i sztucznej inteligencji, ta współpraca powinna umożliwić dalsze zrozumienie zarówno naturalnej, jak i sztucznej inteligencji. Może to pomóc w odpowiedzi na kilka kluczowych pytań dotyczących tego, co to znaczy być człowiekiem i co to znaczy żyć obok sztucznej inteligencji.