Wstęp

Wiele prac związanych ze sztuczną inteligencją opiera się na koncepcjach i teoriach opracowanych przez filozofów i logików. W tej części przedstawiono tę podstawową pracę, omawiając różne koncepcje sztucznej inteligencji, filozoficzne marzenie o mechanizacji ludzkiego rozumowania, koncepcyjne korzenie sztucznej inteligencji i główne teorie umysłu, które leżą u podstaw różnych kierunków badań nad sztuczną inteligencją.

Czym jest sztuczna inteligencja?

To samo w sobie jest głębokim pytaniem filozoficznym, a próby systematycznej odpowiedzi na nie mieszczą się w podstawach sztucznej inteligencji jako bogatego tematu do analiz i debat. Niemniej jednak można udzielić tymczasowej odpowiedzi: sztuczna inteligencja to dziedzina poświęcona tworzeniu artefaktów zdolnych do wykazywania, w kontrolowanym, dobrze rozumianym środowisku i przez dłuższy czas, zachowań, które uważamy za inteligentne, lub, bardziej ogólnie, zachowań, które uważamy, że znajdujemy się w sercu tego, czym jest posiadanie umysłu. Oczywiście odpowiedź ta rodzi dalsze pytania, w szczególności dotyczące tego, co dokładnie stanowi inteligentne zachowanie, co to znaczy mieć umysł i jak ludziom faktycznie udaje się zachowywać inteligentnie. Ostatnie pytanie ma charakter empiryczny; odpowiedź należy do psychologii i nauk kognitywnych. Jest to jednak szczególnie istotne, ponieważ jakikolwiek wgląd w ludzką myśl może pomóc nam w zbudowaniu maszyn działających podobnie. Rzeczywiście, jak okaże się , sztuczna inteligencja i kognitywistyka rozwinęły się równoległymi i ściśle splecionymi ścieżkami; ich historii nie da się opowiedzieć osobno. Drugie pytanie, które pyta, jaki jest znamię tego, co mentalne, ma charakter filozoficzny. Sztuczna inteligencja nadała tej kwestii pilność i i odwrotnie, zobaczymy, że uważna filozoficzna kontemplacja tej kwestii wpłynęła na przebieg samej sztucznej inteligencji. Wreszcie pierwsze wyzwanie polegające na dokładnym określeniu, co należy uznać za inteligentne zachowanie, tradycyjnie stawiano czoła proponowaniu konkretnych testów behawioralnych, których pomyślne zdanie oznaczałoby obecność inteligencji. Najbardziej znanym z nich jest test Turinga (TT), wprowadzony przez Turinga (1950). W TT człowiek i komputer są zamknięte w zamkniętych pomieszczeniach, a ludzki sędzia, nie wiedząc, w którym z dwóch pokoi znajduje się który zawodnik, zadaje im pytania e-mailem (właściwie telegrafem, aby użyć oryginalnego terminu). Jeżeli na podstawie zwróconych odpowiedzi sędzia nie będzie w stanie zrobić nic lepszego niż 50/50, kiedy wydając werdykt, w którym pokoju znajduje się który gracz, mówimy, że dany komputer przeszedł test TT. Zdaniem Turinga, komputer zdolny przejść TT należy uznać za maszynę myślącą. Jego twierdzenie wzbudziło kontrowersje, choć wydaje się niezaprzeczalne, że zachowania językowe wymagane przez TT są rutynowo uznawane za sedno ludzkiego poznania. Część kontrowersji wynika z bezwstydnie behawioralnych założeń testu. Eksperyment myślowy Blocka „Ciotka Bertha" (1981) miał na celu podważenie tych założeń, argumentując, że nie tylko zachowanie organizmu decyduje o tym, czy jest on inteligentny. Musimy także rozważyć, w jaki sposób organizm osiąga inteligencję. Oznacza to, że należy wziąć pod uwagę wewnętrzną organizację funkcjonalną systemu. Był to kluczowy punkt funkcjonalizmu, kolejny ważny nurt filozoficzny sztucznej inteligencji, do którego powrócimy później. Inna krytyka TT dotyczy tego, że jest on nierealistyczny i mógł nawet utrudniać postęp sztucznej inteligencji w zakresie bezcielesnej inteligencji. Jak się przekonamy, wielu myślicieli doszło do wniosku, że bezcielesne artefakty posiadające inteligencję na poziomie ludzkim są mrzonką, praktycznie niemożliwą do zbudowania, jeśli nie wręcz absurdalną koncepcyjnie. W związku z tym Harnad (1991) upiera się, że od artefaktów, które oznaczałyby sukces sztucznej inteligencji, wymagana jest zdolność sensomotoryczna, i proponuje Total TT (TTT) jako ulepszenie w stosunku do TT. Podczas gdy w TT bezcielesny program komputerowy mógłby, przynajmniej w zasadzie, przejść, przechodnie TTT muszą być robotami zdolnymi do działania w środowisku fizycznym w sposób nieodróżnialny od zachowań przejawianych przez wcielone osoby ludzkie poruszające się po świecie

fizycznym. Definiując sztuczną inteligencję jako dziedzinę poświęconą artefaktom inżynieryjnym, które są w stanie przejść testy TT, TTT i różne inne testy, można śmiało powiedzieć, że mamy do czynienia ze słabą sztuczną inteligencją. Inaczej mówiąc, słaba sztuczna inteligencja ma na celu budowanie maszyn, które działają inteligentnie, bez zajmowania stanowiska w sprawie tego, czy maszyny rzeczywiście są inteligentne. Istnieje inna odpowiedź na pytanie „Co to jest sztuczna inteligencja?" pytanie: AI to dziedzina poświęcona budowaniu osób, kropka. Ten rodzaj sztucznej inteligencji to tak zwana silna sztuczna inteligencja, ambitna forma tej dziedziny, którą trafnie podsumował Haugeland:

Podstawowym celem [badań nad sztuczną inteligencją] nie jest jedynie naśladowanie inteligencji lub tworzenie sprytnych podróbek. Zupełnie nie. AI chce tylko prawdziwego artykułu: maszyn z umysłami, w pełnym i dosłownym tego słowa znaczeniu. To nie jest fantastyka naukowa, ale prawdziwa nauka, oparta na koncepcji teoretycznej tak głębokiej, jak śmiałej: mianowicie, że w istocie sami jesteśmy komputerami.

(Haugeland 1985, s. 2)

Ta „teoretyczna koncepcja" ludzkiego umysłu jako komputera stała się podstawą większości dotychczasowych badań nad sztuczną inteligencją. Stała się znana jako obliczeniowa teoria umysłu; wkrótce omówimy to szczegółowo. Z drugiej strony inżynieria sztucznej inteligencji, która sama w sobie opiera się na filozofii, jak w przypadku nieustannej próby mechanizacji rozumowania, omówionej w następnej sekcji, może być realizowana w służbie zarówno słabej, jak i silnej sztucznej inteligencji.

Filozoficzna sztuczna inteligencja: przykład mechanizacji rozumowania

Nie byłoby nierozsądnym opisać klasyczną kognitywistykę jako rozszerzoną próbę zastosowania metod teorii dowodu do modelowania myślenia.

W tej części omówiono obszar będący przykładem AI, który jest powiązany z filozofią (w przeciwieństwie do filozofii AI). Jest to obszar, który powinien przede wszystkim znać każdy student filozofii i sztucznej inteligencji. Częściowo dzieje się tak dlatego, że inne problemy związane ze sztuczną inteligencją, przynajmniej częściowo natury filozoficznej, są ściśle powiązane z próbami zmechanizowania rozumowania na poziomie ludzkim. Arystoteles uważał racjonalność za istotną cechę ludzkiego umysłu. Myśl dedukcyjna, wyrażona za pomocą sylogizmów, była cechą charakterystyczną takiej racjonalności, a także podstawowym narzędziem intelektualnym (organonem) wszelkiej nauki. Być może najgłębszym wkładem Arystotelesa w sztuczną inteligencję była idea formalizmu. Pogląd, że pewne wzorce myślenia są ważne ze względu na ich formę syntaktyczną, niezależnie od treści, był niezwykle potężną innowacją i to właśnie to pojęcie pozostaje w centrum współczesnej obliczeniowej teorii umysłu i co nazwaliśmy powyżej silną sztuczną inteligencją. Ze względu na znaczenie, jakie historycznie przypisywano dedukcji w filozofii (począwszy od Arystotelesa, poprzez Euklidesa, a później Bacona, Hobbesa, Leibniza i innych), sama idea inteligentnej maszyny była często równoznaczna z maszyną, która może wykonywać wnioskowanie logiczne: takie, które pozwala na prawidłowe wyciąganie wniosków z przesłanek. Automatyczne dowodzenie twierdzeń (ATP), jak dziś nazywa się tę dziedzinę, było zatem integralną częścią sztucznej inteligencji od samego początku, chociaż, jak zobaczymy, jego znaczenie było przedmiotem gorących dyskusji, zwłaszcza w ostatnich dziesięcioleciach. Ogólnie rzecz biorąc, problem mechanizacji dedukcji ma co najmniej trzy różne warstwy. W kolejności rosnącej trudności mamy:

* Sprawdzanie dowodu: Biorąc pod uwagę dedukcję D, która ma na celu wyprowadzenie wniosku P z szeregu przesłanek P₁ &heip;P<sub>n, zdecyduj, czy czy nie. D jest poprawne.

* Odkrycie dowodu: Biorąc pod uwagę szereg przesłanek P1,…, Pn i domniemany wniosek P, oceń, czy P wynika logicznie z przesłanek, a jeśli tak, to przedstawić ich formalne wyprowadzenie.

* Tworzenie hipotez: Biorąc pod uwagę szereg przesłanek P₁ …P_n, wygeneruj „interesujący” wniosek P, który prawdopodobnie logicznie wynika z przesłanek.

Z technicznego punktu widzenia pierwszy problem jest najłatwiejszy. W przypadku logiki predykatów z równością problem sprawdzenia poprawności danej dedukcji jest nie tylko algorytmicznie rozwiązywalny, ale i skuteczny. Niemniej jednak problem ten jest pełen interesujących kwestii filozoficznych i technicznych, a jego znaczenie dla sztucznej inteligencji wcześnie zdał sobie sprawę McCarthy (1962), który napisał, że „sprawdzanie dowodów matematycznych jest potencjalnie jednym z najciekawszych i najbardziej przydatnych zastosowań komputerów automatycznych”. Drugi problem jest znacznie trudniejszy. Wczesne wyniki teorii funkcji rekurencyjnych wykazały, że nie ma maszyny Turinga, która mogłaby zdecydować, czy dowolny wzór logiki pierwszego rzędu jest poprawny (był to problem Entscheidungsproblem Hilberta). Zatem z tezy Churcha wynika, że problem jest algorytmicznie nierozwiązywalny – nie ma ogólnej metody mechanicznej, która zawsze podejmie właściwą decyzję w skończonym czasie. Jednak ludzie również nie mają gwarancji, że zawsze rozwiążą problem (i często tego nie robią). W związku z tym sztuczna inteligencja może szukać konserwatywnych przybliżeń, które dobrze sprawdzają się w praktyce: programów, które podają właściwą odpowiedź tak często, jak to możliwe, a w przeciwnym razie w ogóle nie dają odpowiedzi (albo wyraźnie zawodzą, albo działają w nieskończoność, dopóki ich nie zatrzymamy). Problem został rozwiązany wcześnie w przypadku słabszych formalizmów z pozornie obiecującymi wynikami: Teoretyk logiki (LT) Newella, Simona i Shawa, zaprezentowany na inauguracyjnej konferencji AI w Dartmouth w 1956 r., zdołał udowodnić trzydzieści osiem z pięćdziesięciu dwóch twierdzenia zdaniowo-logiczne Principia Mathematica. Inne godne uwagi wczesne wysiłki obejmowały wdrożenie arytmetyki Presburgera przez Martina Davisa w 1954 r. w Instytucie Studiów Zaawansowanych w Princeton (Davis 2001), procedurę Davisa-Putnama, której odmiany są dziś używane w wielu dowodach opartych na spełnialności oraz imponujący system logiki pierwszego rzędu zbudowany przez Wanga (1960). Należy zauważyć, że podczas gdy LT został celowo zaprojektowany do symulowania ludzkiego rozumowania i procesów rozwiązywania problemów, autorzy tych innych systemów uważali, że naśladowanie ludzkich procesów jest niepotrzebnie ograniczające i że lepsze wyniki można osiągnąć, rezygnując z wiarygodności poznawczej . Był to wczesny przejaw napięcia, które nadal jest odczuwalne w tej dziedzinie i które odpowiada rozróżnieniu między silnymi i słabymi formami sztucznej inteligencji: sztuczna inteligencja jako nauka, w szczególności jako badanie ludzkiej myśli, kontra sztuczna inteligencja jako inżynieria – budowa inteligentnych systemy, których działanie nie musi przypominać wewnętrznego działania ludzkiego ognia. Odkrycie Robinsona dotyczące unifikacji i metody rozdzielczości (Robinson 1965) zapewniło ogromny impuls w tej dziedzinie. Większość współczesnych automatycznych dowodów twierdzeń opiera się na rozdzielczości. Inne ważne formalizmy obejmują obrazy semantyczne i logikę równań . Chociaż w ciągu ostatnich dziesięciu lat nastąpił imponujący postęp, obecnie najbardziej wyrafinowane ATP nadal są kruche i czasami zawodzą w przypadku problemów, które byłyby trywialne dla studentów. Trzeci problem, dotyczący generowania przypuszczeń, jest najtrudniejszy, ale także najciekawszy. W końcu domysły nie spadają z nieba. Mając do dyspozycji mnóstwo informacji, ludzie – szczególnie matematycy – regularnie przedstawiają interesujące przypuszczenia, a następnie starają się je udowodnić, często z sukcesem. Ten proces odkrywania (wraz z tworzeniem nowych koncepcji) jest jednym z najbardziej twórczych działań ludzkiego intelektu. Sama trudność w obliczeniowym symulowaniu tej kreatywności jest z pewnością głównym powodem, dla którego sztuczna inteligencja poczyniła w tym zakresie niewielkie postępy. Innym powodem jest jednak to, że przez większą część poprzedniego stulecia (a właściwie począwszy od Fregego w XIX wieku) logicy i filozofowie zajmowali się prawie wyłącznie uzasadnianiem, a nie

odkrywaniem. Dotyczyło to nie tylko rozumowania dedukcyjnego, ale także rozumowania indukcyjnego, a nawet ogólnie teorii naukowej (Reichenbach 1938). Panowało powszechne przekonanie, że procesem odkrywania powinni zajmować się psychologowie, a nie filozofowie i logicy. Co ciekawe, nie miało to miejsca przed Fregem. Filozofowie tacy jak Kartezjusz (1988), Bacon (2002), Mill (1874) i Peirce (1960) próbowali racjonalnie zbadać proces odkrywania i sformułować zasady nim kierujące. Począwszy od Hansona (1958) w nauce i Lakatosa (1976) w matematyce (na tego ostatniego duży wpływ miał Pólya (1945)) filozofowie zaczęli ponownie kłaść nacisk na odkrycia. Badacze zajmujący się sztuczną inteligencją próbowali także obliczeniowo modelować odkrywanie, zarówno w naukach ścisłych , jak i matematyce (Lenat 1976, 1983), a ten kierunek prac doprowadził do innowacji w zakresie sztucznej inteligencji w zakresie uczenia maszynowego, takich jak programowanie genetyczne (Koza 1992) i programowanie w logice indukcyjnej . Jednak sukcesy były ograniczone, a podstawowe zastrzeżenia filozoficzne wobec algorytmicznego podejścia do odkrywania i kreatywności w ogóle – na przykład wysunięte przez Hempla (1985) – pozostają ostre. Głównym problemem jest pozornie holistyczny charakter wyższych procesów poznawczych, takich jak twórcze rozumowanie, oraz trudność w sformułowaniu rygorystycznej charakterystyki istotności. Bez precyzyjnego pojęcia relewancji, które można zastosować w obliczeniach, wydaje się, że niewiele jest nadziei na postęp w rozwiązaniu problemu generowania wniosków lub któregokolwiek innego podobnego problemu, w tym generowania koncepcji i tworzenia hipotez abdukcyjnych. W obliczu stosunkowo skromnego postępu w zakresie trudnych problemów rozumowania i pod wpływem różnych krytyk symbolicznej sztucznej inteligencji (patrz sekcja 2.6), niektórzy badacze sztucznej inteligencji przypuścili poważne ataki na logikę formalną, krytykując ją jako nadmiernie sztywny system, który nie zapewnia dobrego modelu wybitnie elastycznych mechanizmów ludzkiego rozumowania. W związku z tym próbowali odwrócić uwagę i wysiłki tej dziedziny od rygorystycznego rozumowania dedukcyjnego i indukcyjnego, kierując je zamiast tego w stronę „rozumowania zdroworozsądkowego". Wiele prac poświęcono opracowaniu formalnych systemów modelowania zdroworozsądkowego rozumowania (Davis i Morgenstern 2004). Jednak krytycy zarzucają, że takie wysiłki nie uwzględniają najważniejszego celu. Na przykład Winograd pisze,

że Minsky za niepowodzenie w wyjaśnianiu zwykłego rozumowania obwinia sztywność logiki, a nie stawia bardziej fundamentalnych pytań o naturę wszelkich reprezentacji symbolicznych i systemów formalnych (choć być może nielogicznych) zasad manipulacji nimi. Istnieją podstawowe ograniczenia tego, co można zrobić z manipulacją symbolami, niezależnie od tego, ile „różnych, użytecznych sposobów łączenia rzeczy w całość" ktoś wymyśli. Redukcja umysłu do pozbawionych kontekstu fragmentów jest ostatecznie niemożliwa i wprowadza w błąd. (Winograd 1990, s. 172)

Jak zobaczymy w dalszej części, podobną krytykę wysunęli Dreyfus (1992) i inni, którzy argumentowali, że manipulacja symbolami nie może wyjaśniać tak podstawowych cech ludzkich, jak intuicja, osąd i wyobraźnia, z których wszystkie mogą odgrywać kluczową rolę. ogólnie w zakresie wnioskowania i rozwiązywania problemów; oraz że ludzkiemu rozumowaniu nigdy nie uda się dorównać żadnemu pozbawionemu kontekstu i bezcielesnemu (lub „niesytuowanemu") systemowi, który działa poprzez formalne reprezentowanie informacji symbolicznej i manipulowanie nią.

Historyczne i koncepcyjne korzenie AI

Oficjalne początki sztucznej inteligencji miały miejsce w 1956 r. podczas małej, ale obecnie słynnej letniej konferencji w Dartmouth College w Hanowerze w stanie New Hampshire. (Obchody pięćdziesiątej rocznicy tej konferencji AI@50 odbyły się w lipcu 2006 roku w Dartmouth i wróciło pięciu pierwotnych uczestników. Część tego, co wydarzyło się podczas tej historycznej konferencji, opisano w ostatniej części tej sekcji). Uczestniczyło w nim dziesięciu myślicieli, w tym John McCarthy (który pracował w Dartmouth w 1956 r.), Claude Shannon, Marvin Minsky, Arthur Samuel, Trenchard Moore

(najwyraźniej najmłodszy uczestnik i jedyny sporządzający notatki na pierwotnej konferencji), Ray Solomonoff, Oliver Selfridge, Allen Newell i Herbert Simon. Z miejsca, w którym stoimy teraz, kilka lat po rozpoczęciu nowego tysiąclecia, konferencja w Dartmouth zapada w pamięć z wielu powodów, w tym z tej pary: (1) ukuto tam termin „sztuczna inteligencja" (i od dawna jest mocno zakorzeniony, pomimo niechęci do niego) tego dnia przez niektórych uczestników, np. Moore'a); (2) Newell i Simon ujawnili, że program – Logic Theorist (LT) – uzgodniony przez uczestników konferencji (i w rzeczywistości przez prawie wszystkich, którzy dowiedzieli się o nim wkrótce po wydarzeniu w Dartmouth) jest niezwykłym osiągnięciem. Jak już wspomnieliśmy, metoda LT była w stanie udowodnić elementarne twierdzenia rachunku zdań i została uznana za niezwykły krok w kierunku przedstawienia rozumowania na poziomie ludzkim w konkretnych obliczeniach. Jednak z filozoficznego punktu widzenia ani konferencja z 1956 r., ani wspomniany artykuł Turinga Minda z 1950 r. nie są bliskie wyznaczenia początku sztucznej inteligencji. Hobbes przewidywał silną sztuczną inteligencję już w XVII wieku, kiedy w słynny sposób ogłosił, że „racjonalizacja to obliczanie". Mniej więcej w tej samej epoce Leibniz marzył o „rachunku uniwersalnym", w którym wszystkie spory można byłoby rozstrzygać na podstawie obliczeń na pamięć. Kartezjusz rozważał już coś w rodzaju testu Turinga na długo przed Turingiem, aczkolwiek przyjął raczej pesymistyczne podejście do tej kwestii, być może w nieco wyrafinowany sposób:

[Gdyby] istniały maszyny, które byłyby podobne do naszego ciała i naśladowałyby nasze działania, o ile było to moralnie możliwe, zawsze mielibyśmy dwa bardzo pewne testy, dzięki którym moglibyśmy rozpoznać, że mimo wszystko nie były one prawdziwi mężczyźni. Po pierwsze, nigdy nie mogliby używać mowy ani innych znaków tak jak my, gdy zapisujemy nasze myśli dla dobra innych. Łatwo bowiem zrozumieć budowę maszyny, która może wypowiadać słowa, a nawet emitować na nią pewne reakcje na działania o charakterze cielesnym… Ale nigdy nie zdarza się, aby na różne sposoby układała ona swoją mowę, aby na wszystko odpowiednio odpowiedzieć można to powiedzieć w jego obecności, jak może to zrobić nawet najniższy typ człowieka. Druga różnica polega na tym, że chociaż maszyny mogą wykonywać pewne rzeczy równie dobrze lub może lepiej niż ktokolwiek z nas, w innych z pewnością zawodzą, w ten sposób możemy odkryć, że nie działały na podstawie wiedzy, a jedynie od rozmieszczenia ich organów. (Kartezjusz 1911, s. 116)

Ale choć uroczystą inauguracją sztucznej inteligencji była konferencja w Dartmouth w 1956 r., a filozofowie od stuleci rozmyślali o maszynach i inteligencji, kluczowe koncepcyjne początki sztucznej inteligencji można znaleźć na skrzyżowaniu dwóch najważniejszych osiągnięć intelektualnych XX wieku:

* „rewolucja poznawcza"4, która rozpoczęła się w połowie lat pięćdziesiątych XX wieku i która obaliła behawioryzm i zrehabilitowała psychologię mentalistyczną;

* teoria obliczalności rozwijana w ciągu ostatnich kilku dekad przez pionierów, takich jak Turing, Church, Kleene i Gödel.

Znaczenie każdego z nich dla sztucznej inteligencji zostanie pokrótce omówione poniżej. Rewolucję poznawczą kojarzy się zazwyczaj z pracami George'a Millera i Noama Chomsky'ego z lat pięćdziesiątych XX wieku, zwłaszcza z osławioną recenzją teorii języka Skinnera przez tego ostatniego (Chomsky 1996). Przewidywali to już w latach czterdziestych XX wieku McCulloch i Pitts (1943) oraz inni pionierzy cybernetyki, którzy już zwracali uwagę na podobieństwa między ludzkim myśleniem a przetwarzaniem informacji, a także na podstawie wyników eksperymentów uzyskanych przez psychologów, takich jak Tolman (1948), który badając nawigację w labiryncie przez szczury, przedstawił dowody na istnienie „map poznawczych". Szczególnie wpływowy był słynny argument Chomsky'ego dotyczący „ubóstwa bodźców", mówiący, że efektywności i szybkości przyswajania języka w dzieciństwie nie można wytłumaczyć wyłącznie odwoływaniem się do skąpych danych, z którymi dzieci mają kontakt we

wczesnych latach życia; raczej zmuszają do postulowania wrodzonych reguł i reprezentacji mentalnych, które kodują kompetencje językowe. Mocne dowody na istnienie reprezentacji mentalnych dostarczyły także ustalenia eksperymentalne dotyczące pamięci, takie jak wyniki Sperlinga (1960), które wykazały, że ludzie zazwyczaj przechowują więcej informacji, niż są w stanie opisać. W końcu pamięć dostarcza być może najjaśniejszego przypadku reprezentacji umysłowej; absurdem wydaje się zaprzeczanie, że ludzie przechowują informacje, to znaczy, że mamy jakąś wewnętrzną reprezentację informacji, takich jak rok urodzenia lub imiona naszych rodziców. To wszystko jest zdroworozsądkowe aż do trywialności, podobnie jak twierdzenie, że ludzie na co dzień mówią tak, jakby naprawdę mieli przekonania, nadzieje, pragnienia itd.; i rzeczywiście większość behawiorystów nie zaprzeczyłaby tym twierdzeniom. Zaprzeczali teoretycznej legitymizacji wyjaśniania ludzkich zachowań poprzez twierdzenie, że nieobserwowalne byty mentalne (takie jak wspomnienia) lub że terminologia intencjonalna ma jakiekolwiek miejsce w nauce o umyśle. Zasadniczo behawioryzm, będący doktryną pozytywistyczną, cechował się nieufnością do wszystkiego, czego nie można bezpośrednio zaobserwować i ogólną niechęcią do teorii. Był to paradygmat dominujący w psychologii przez większą część XX wieku, aż do połowy lat pięćdziesiątych XX wieku, aż w końcu został zdetronizowany przez nowe podejście „poznawcze".

Obliczeniowe teorie umysłu i problem treści umysłowych

Kiedy już poczyniono pierwsze kroki i otwarcie dopuszczono reprezentacje mentalne w naukowym teoretyzowaniu na temat umysłu, „metafora komputerowa" – z którą flirtowali już badacze tacy jak Newell i Simon – dojrzała do eksplozji. Wiadomo przecież, że komputery przechowują w pamięci uporządkowane dane i rozwiązują problemy poprzez systematyczne manipulowanie tymi danymi i wykonywanie odpowiednich instrukcji. Być może podobny model mógłby wyjaśnić – i ostatecznie pomóc w odtworzeniu – ludzkiej myśli. W istocie postulowanie reprezentacji mentalnych samo w sobie nie byłoby zbyt daleko idące, gdyby ich skuteczności przyczynowej nie można było wyjaśnić w sposób mechanistyczny i systematyczny. To prawda, że ustrukturyzowane reprezentacje mentalne są niezbędne do poznania wyższego rzędu; ale w jaki sposób takie reprezentacje faktycznie powodują racjonalne myślenie i działanie? Teoria obliczeń została powołana właśnie po to, aby sprostać tej ważnej potrzebie teoretycznej. Rezultat stał się znany jako obliczeniowa teoria umysłu (w skrócie CTM), doktryna nierozerwalnie powiązana z silną sztuczną inteligencją. W następnym akapicie pokrótce omówimy główne założenia CTM. Pierwszą podstawową ideą CTM jest wyjaśnienie intencjonalnych stanów psychicznych poprzez obliczenie analizy Russella (1940) zdań intencjonalnych, takich jak „Tom wierzy, że 7 jest liczbą pierwszą", jako postaw zdaniowych obejmujących postawę psychologiczną A (w tym wiara w przypadek) w kierunku zdania P (w tym przypadku, że 7 jest liczbą pierwszą). Dokładniej, bycie w stanie psychicznym obejmującym postawę A i zdanie P oznacza bycie w pewnym związku R_A z reprezentacją mentalną M_P, której znaczenie wynosi P. Upraszczając, posiadanie przekonania, że 7 jest liczbą pierwszą, jest mieć mentalną reprezentację w swoim „skrzynce przekonań", co oznacza, że 7 jest liczbą pierwszą. Samo przedstawienie jest symboliczne. Oznacza to, że twoje „skrzynka przekonań" zawiera symbol struktury symbolicznej, której znaczenie (lub „treść") jest takie, że 7 jest liczbą pierwszą. Zatem reprezentacje mentalne mają zarówno składnię, jak i semantykę, podobnie jak zdania języków naturalnych. Stanowią, że tak powiem, „język myśli", czyli mentalese. Ale to tylko ich składnia – składnia, która ostatecznie sprowadza się do kształtu fizycznego – czyni je przyczynowo skutecznymi. Jest to wiarygodna historia, ponieważ, jak pokazały prace nad logiką i obliczalnością, istnieją czysto syntaktyczne transformacje struktur symbolicznych, które są jednak wrażliwe na semantykę. Być może najlepszym przykładem są dowody dedukcyjne: manipulując wzorami wyłącznie na podstawie ich właściwości składniowych, można wydobyć z nich inne wzory, które logicznie z nich wynikają. Składnia może zatem odzwierciedlać semantykę lub, jak to ujął Haugeland , „jeśli zadbasz o składnię, semantyka zadba o siebie sama". W tym modelu proces mentalny to sekwencja symboli reprezentacji mentalnych,

które wyrażają treść zdań odpowiednich myśli. Przyczyny i skutki każdej reprezentacji mentalnej oraz to, co faktycznie może ona zrobić, są określone przez jej składnię „w podobny sposób, w jaki geometria klucza określa, które zamki zostaną otwarte” . Cały proces jest zarządzany przez algorytm – zestaw instrukcji określających, w jaki sposób reprezentacje następują po sobie w ogólnym toku myślenia. To jest druga podstawowa idea CTM. Umysł jest zatem postrzegany jako „silnik syntaktyczny” napędzający silnik semantyczny i, przynajmniej w zasadzie, jego działanie można odtworzyć na komputerze. Naturalnym rozwinięciem CTM jest funkcjonalizm maszyny Turinga, o którym po raz pierwszy wspomniał Putnam (1960) w wpływowym artykule, który pomógł popchnąć rewolucję poznawczą (przynajmniej w kręgach filozoficznych), podważyć behawioryzm i ukształtować światopogląd. silna sztuczna inteligencja. Funkcjonalizm w ogóle to pogląd, że istoty stanu psychicznego nie można znaleźć w biologii mózgu (lub w fizyce, która leży u podstaw sprzętu jego jednostki centralnej, w przypadku maszyny), ale raczej w roli, jaką stan odgrywa w życiu psychicznym (lub obliczeniach), a zwłaszcza w związkach przyczynowych, jakie wywiera na bodźce (wejścia), zachowanie (wyjścia) i inne stany mentalne (obliczeniowe). W szczególności funkcjonalizm maszyny Turinga to idea, zgodnie z którą umysł jest w istocie gigantyczną maszyną Turinga, której działanie jest określone przez zestaw instrukcji mówiących, że jeśli umysł znajduje się w pewnym stanie s i otrzymuje określony sygnał wejściowy x, przejście następuje doprowadzany do stanu s’ i emitowany jest sygnał wyjściowy y. Najpopularniejsze – i najmniej nieprawdopodobne – wersje funkcjonalizmu maszyny Turinga pozwalają na przejścia probabilistyczne. Również ściśle powiązana z CTM (w rzeczywistości silniejsza od niej) jest hipoteza systemu symboli fizycznych (PSSH) wysunięta przez Newella i Simona (1976). Zgodnie z nią fizyczny system symboli „posiada niezbędne i wystarczające środki do ogólnego inteligentnego działania”, podczas gdy fizyczny system symboli jest „maszyną, która z biegiem czasu wytwarza ewoluujący zbiór struktur symboli” – struktura symboli będąca zbiorem żetonów symboli „powiązanych w jakiś fizyczny sposób (np. jeden token znajdujący się obok drugiego)” i podlegających różnorodnym operacjom składniowym, w szczególności „tworzeniu, modyfikacji, reprodukcji i zniszczenie.” Newell i Simon uważali maszyny wykonujące programy do przetwarzania list typu LISP za prototypowe przykłady fizycznych systemów symboli. Chociaż istniały różne wewnętrzne spory (np. dotyczące kwestii wrodzonej), w tej czy innej formie CTM, PSSH i funkcjonalizm maszyny Turinga razem luźno charakteryzują „klasyczną” lub „symboliczną” sztuczną inteligencję, czyli to, co Haugeland nazwał GOFAI („dobra, staromodna sztuczna inteligencja”). Wszystkie trzy zostały przyjęte jako merytoryczne tezy empiryczne, CTM i funkcjonalizm maszyny Turinga na temat ludzkiego umysłu oraz PSSH na temat inteligencji w ogóle (również GOFAI zostało wyraźnie scharakteryzowane przez Haugelanda jako empiryczna doktryna nauk kognitywnych). Wyznaczają parametry i cele większości badań nad sztuczną inteligencją przez co najmniej pierwsze trzy dekady tej dziedziny i nadal wywierają dominujący wpływ, chociaż, jak zobaczymy, nie są już jedyną grą w mieście, ponieważ poniosły znaczne straty niepowodzeń w wyniku silnych ataków, które wywołały poważne problemy koncepcyjne i empiryczne związane z podejściem GOFAI. Według CTM złożone myśli są reprezentowane przez złożone struktury symboliczne w podobny sposób, w jaki w językach naturalnych i logikach formalnych złożone zdania są rekurencyjnie budowane z prostszych elementów. Zatem mentalna reprezentacja złożonej myśli, takiej jak „Wszyscy ludzie są śmiertelni”, zawiera składowe mentalne reprezentacje takich pojęć, jak „śmiertelnik” i „ludzie”, a także „wszyscy” i „są”. Składniki te są w jakiś sposób składane razem (i ostatecznie nauka powinna być w stanie określić szczegóły tego, jak takie symboliczne operacje przeprowadzane są w mózgu), tworząc złożoną myśl, której treścią jest to, że wszyscy ludzie są śmiertelni. W ten sposób złożone myśli uzyskują swoje znaczenie, łącząc znaczenia ich składników. Tego rodzaju opowieść kompozycyjna – podobna do semantyki kompozycyjnej, za którą opowiadają się Fregego i Tarski – jest możliwa tylko wtedy, gdy istnieje zbiór prymitywów, które można wykorzystać jako ostateczne elementy składowe bardziej złożonych reprezentacji. Centralnym pytaniem dla CTM, które ma bezpośredni odpowiednik w sztucznej inteligencji, jest to, w jaki sposób

te prymitywy nabierają znaczenia. Mówiąc dokładniej, pytanie brzmi, w jaki sposób prymitywne umysły znajdujące się w naszych mózgach (lub w centralnej jednostce przetwarzającej robota) radzą sobie z obiektami i stanami rzeczy poza naszymi mózgami – przedmiotami, które mogą nawet nie istnieć i stanami rzeczy, które mogą nawet nie zaistnieć. Nazywa się to również problemem uziemienia symboli (Harnad 1990). Nie jest to jedynie filozoficzna zagadka dotycząca ludzkiego umysłu, ani nawet protonaukowe pytanie z zakresu psychologii. Ma to bezpośrednie implikacje inżynieryjne dla sztucznej inteligencji, ponieważ wiarygodna odpowiedź na to pytanie może przełożyć się na metodologię budowy robota, która potencjalnie pozwoli uniknąć niektórych z najbardziej niszczycielskich zastrzeżeń wobec CTM (zostaną one omówione w następnej sekcji). Taki robot „myśliłby", wykonując obliczenia na formalnych strukturach symbolicznych (jak przypuszczamy według CTM), ale mimo to byłby na tyle osadzony w świecie rzeczywistym, że można by powiedzieć, że osiąga zrozumienie pozasymboliczne (tak jak my). . Oczywiście nie można sugerować, że ewolucja obdarzyła nas wszystkimi właściwymi, prymitywnymi symbolami, posiadającymi wbudowane wszystkie właściwe znaczenia, ponieważ ewolucja nie mogła przewidzieć takich rzeczy jak termostaty czy satelity. W odpowiedzi wyjaśniono wiele teorii, a wszystkie mieszczą się pod hasłem „naturalizującej treści", „naturalizującej semantyki" lub „naturalizującej intencjonalności". Celem jest dostarczenie fizykalistycznego opisu tego, jak mentalne symbole w naszych głowach potrafią odnosić się do rzeczy zewnętrznych w stosunku do nas (lub, ujmując kwestię niezależnie od CTM, w jaki sposób stany mentalne mogą nabrać znaczenia). Innymi słowy, szukany typ relacji jest redukcyjny i materialistyczny; należy to wyrazić w nieintencjonalnym słownictwie czystych nauk fizycznych. Poniżej dokonamy krótkiego przeglądu trzech najważniejszych prób przedstawienia naturalizowanych opisów znaczenia: teorii informacyjnych, teorii ewolucyjnych i semantyki ról pojęciowych. Istotą teorii informacji jest pojęcie kowariancji. Pomysł jest taki, że jeśli wielkość x jest systematycznie zmienna z wielkością y, to x niesie informację o y. Prędkościomierz samochodu systematycznie zmienia się wraz z prędkością samochodu i w ten sposób przekazuje informacje na jej temat. W związku z tym możemy postrzegać prędkościomierz jako system zamierzony, ponieważ jego odczyty dotyczą prędkości samochodu. Podobnie mówimy, że dym „oznacza" ogień, a dym ten niesie informację o ogniu. Takie właśnie znaczenie słowa „oznacza" Grice nazwał znaczeniem naturalnym, co było zapowiedzią teorii semantyki informacyjnej. Ponownie, ponieważ dym i ogień są nomologicznie kowariantne, możemy powiedzieć, że jedno dotyczy drugiego. Mamy tu zatem początki naturalizowanego podejścia do znaczenia, które traktuje intencjonalność jako powszechne zjawisko naturalne, a nie zjawisko specyficznie mentalne. Jeśli chodzi o semantykę mentalną, sedno takich teorii – mówiąc nieco upraszczając – jest takie, że znaczenie symbolu jest zdeterminowane przez jakiekolwiek symbole tego symbolu, symbol systematycznie (nomologicznie) współzmienny z. Jeśli symbol pewnego mentalnego symbolu H pojawia się w naszym mózgu za każdym razem, gdy pojawia się przed nami koń, wówczas H niesie informację o koniach, a zatem oznacza konia. Teorie ewolucji utrzymują z grubsza, że stany intencjonalne są adaptacjami w taki sam sposób, w jaki wątroba i kciuki są adaptacjami, i że treść (znaczenie) stanu intencjonalnego jest funkcją, dla której został wybrany, to znaczy celem, dla którego służy. Jak wszystkie teorie adaptacjonistyczne, również i ta może zostać oskarżona o panglosjanizm (Gould i Lewontin 1979). Niemniej jednak podstawowa historia nie jest nieprawdopodobna ze względu na, powiedzmy, przekonania o tym, że w pobliżu znajduje się drapieżnik lub chęć zdobycia i zjedzenia bananów pojawiających się w polu widzenia. Treścią takiego przekonania byłoby coś w rodzaju „pod tamtym drzewem jest tygrys", co prawdopodobnie byłoby funkcją, dla której takie przekonania zostały wybrane (aby korelowały z tygrysami pod drzewami), a treść takiego pragnienia brzmiałoby: „Chcę zjeść tamte banany", co ponownie pokrywałoby się z celem, któremu służą takie pragnienia (zdobycie pożywienia i przeżycie). Semantyka ról pojęciowych (CRS) czerpie inspirację ze słynnej teorii znaczenia „używania" Wittgensteina, zgodnie z którą znaczenie elementu językowego (wyrażenia, zdania itp.) to sposób, w jaki element ten jest używany przez osoby posługujące się danym językiem. język. Główną

tezą CRS jest to, że znaczenie mentalnego symbolu S jest ustalone przez rolę, jaką S odgrywa w życiu poznawczym jednostki, a zwłaszcza przez relacje, jakie łączy z innymi symbolami, percepcją i działaniem. Jest zatem bardzo podobny do funkcjonalizmu dotyczącego stanów psychicznych. Łączniki logiczne, takie jak i dostarczają standardowych ilustracji dla teorii użycia znaczenia językowego, a także znaczenia symbolu mentalnego równoważnego roli, jaką odgrywa on w naszych głowach, na przykład zbioru wniosków mentalnych, w których uczestniczy. Teoria wykazuje zauważalne podobieństwa do semantyki operacyjnej języków programowania w informatyce teoretycznej. W kognitywistyce CRS jest znane jako semantyka proceduralna. Popierał ją przede wszystkim Johnson-Laird (1977), a ostro krytykował Fodor. Wiąże się z tym całe zagadnienie eksternalizmu. CRS czyni znaczenie symbolu sprawą na wskroś wewnętrzną, uzależnioną jedynie od relacji, jakie łączy on z innymi symbolami i stanami (w tym stanami percepcyjnymi i behawioralnymi). Ale najwyraźniej kryje się za tym coś więcej. Przynajmniej na pierwszy rzut oka znaczenie wydaje się łączyć symbole ze światem, a nie tylko z innymi symbolami. Znaczenie terminu pies musi mieć coś wspólnego z jego odniesieniem, to znaczy z rzeczywistymi psami, a znaczenie Arystotelesa musi mieć coś wspólnego z rzeczywistym Arystotelesem. Bardziej wyrafinowane wyzwania eksternalistyczne zostały przedstawione przez Putnama i Burge'a, argumentując odpowiednio, że znaczenie jest funkcją ogólnego środowiska fizycznego i społecznego, w którym człowiek jest osadzony. W odpowiedzi opracowano tak zwane dwuczynnikowe CRS, próbując rozróżnić wąskie i szerokie treści mentalne. Wąskie znaczenie tkwi „w głowie" i nie zależy od otaczających go okoliczności, natomiast szerokie znaczenie opiera się na odniesieniu i ma warunki prawdziwości. Wokół tego rozróżnienia wyrosła cała branża i nie mamy tu miejsca, aby się w to zagłębiać.

Nawiasem mówiąc, to właśnie brak połączenia między mentalese a światem (lub między symbolami komputera a światem) był głównym zarzutem Fodora (1978), gdy argumentował przeciwko „semantyce proceduralnej" w stylu sztucznej inteligencji, propagowanej przez Johnsona -Dziedzic. Ten ostatni napisał, że „sztuczne języki używane do przekazywania programów zawierających instrukcje komputerom mają zarówno składnię, jak i semantykę. Ich składnia składa się z reguł pisania dobrze sformułowanych programów, które komputer może interpretować i wykonywać. Ich semantyka składa się z procedur, które komputer ma wykonać"

(Johnson-Laird 1977).

W ważnej krytyce artykułu Johnsona-Lairda, który w pewnym sensie zapowiadał argument Searle'a dotyczący chińskiego pokoju, Fodor zaprotestował, mówiąc, że:

Modele komputerowe nie dostarczają żadnej teorii semantycznej, jeśli przez teorię semantyczną rozumie się relację między językiem a światem. W szczególności semantyka proceduralna nie zastępuje semantyki klasycznej, a jedynie nasuwa pytania na które klasyczni semantycy postanowili odpowiedzieć.

…;

maszyna może skompilować pytanie „Czy Lucy przyniosła deser?" i nie ma zielonego pojęcia, że zdanie dotyczy tego, czy Lucy przyniosła deser. (Fodor 1978,)

Zagadnienia filozoficzne

Trzy główne filozoficzne uwagi krytyczne wobec silnej sztucznej inteligencji, które pomogły zmienić bieg społeczności AI i wskazać nowe kierunki badań, są następujące:

1 Krytyka Huberta Dreyfusa;

2 Krytyka funkcjonalizmu maszyn przez Blocka za pośrednictwem chińskich eksperymentów myślowych;

3 Eksperyment myślowy „Pokój chiński" Searle'a.

Cała trójka pojawiła się w odstępie dziesięciu lat. Wcześniej pojawiło się kilka innych filozoficznych uwag krytycznych na temat silnej sztucznej inteligencji, a od tego czasu pojawiły się kolejne. Ale te trzy wywołały najwięcej dyskusji i wywarły największy wpływ. Pierwsza była krytyka Dreyfusa (Dreyfus 1972). Była to mieszanina argumentów empirycznych i filozoficznych. Z empirycznego punktu widzenia jego głównym zarzutem było to, że badaczom sztucznej inteligencji po prostu nie udało się dostarczyć towarów. Pomimo niezwykle optymistycznych – i często imponujących – wczesnych prognoz, nie udało im się zbudować inteligentnych systemów ogólnego przeznaczenia. Ta linia krytyki została ogólnie odrzucona jako niesłuszna i niesprawiedliwa: nieważna, ponieważ w najlepszym przypadku pokazała, że sztuczna inteligencja jeszcze nie odniosła sukcesu, a nie że nie może kiedykolwiek odnieść sukcesu; i niesprawiedliwe, ponieważ sztuczna inteligencja jest bardzo młodą dziedziną i nie można było oczekiwać rewolucyjnych przełomów technologicznych od dziedziny w jej powijakach, pomimo zbyt entuzjastycznych deklaracji niektórych jej pionierów. Z filozoficznego punktu widzenia Dreyfus argumentował, że nasza zdolność rozumienia świata i innych ludzi jest niedeklaratywnym rodzajem umiejętności know-how, której nie można poddać kodyfikacji zdań w stylu GOFAI. Jest nieartykułowany, przedkoncepcyjny i ma niezbędny wymiar fenomenologiczny, którego nie może uchwycić żaden system oparty na regułach. Dreyfus podkreślił także znaczenie takich zdolności, jak wyobraźnia, tolerancja dwuznaczności i posługiwanie się metaforą, a także zjawisk takich jak świadomość poboczna i percepcja gestalt, z których wszystkie były – i nadal są – odporne na leczenie obliczeniowe. Co najważniejsze, naszym zdaniem Dreyfus podkreślił znaczenie trafności, podkreślając zdolność ludzi do odróżnienia tego, co istotne od tego, co nieistotne i bez wysiłku czerpania z odpowiednich aspektów swojego doświadczenia i wiedzy zgodnie z wymogami ich obecnej sytuacji, jak wymagane przez ich ciągłe zaangażowanie w świat. Słusznie uważał, że przekazanie tych samych możliwości komputerowi cyfrowemu byłoby główną przeszkodą dla sztucznej inteligencji – co nazwał problemem „kontekstu holistycznego". Naszym zdaniem problem trafności pozostaje kluczowym wyzwaniem technicznym dla sztucznej inteligencji, zarówno mocnej, jak i słabej, a także dla obliczeniowej kognitywistyki. Twierdzenie, że ludzie nie wykonują swoich codziennych czynności, przestrzegając zasad, wskazuje na obawę, która jest powracającym problemem w przypadku silnej sztucznej inteligencji i CTM, a nawet ogólnych teorii mentalistycznych, takich jak lingwistyka generatywna Chomsky'ego, i zasługuje na krótkie omówienie w tym miejscu, zanim przejdziemy do przejdź do eksperymentu myślowego Blocka. Wielu filozofów, od Wittgensteina i Quine'a po Dreyfusa, Searle'a i innych, wysunęło ten zarzut pod nieco inną postacią. Ma to związek z tak zwaną psychologiczną rzeczywistością opartych na regułach wyjaśnień poznania, a zwłaszcza z komputerowymi symulacjami procesów umysłowych. Kwestia ta opiera się na rozróżnieniu między opisem i przyczynowością, a także przewidywaniem i wyjaśnianiem. Zbiór reguł (lub tym bardziej program komputerowy) może adekwatnie opisywać zjawisko poznawcze, w tym sensie, że reguły mogą stanowić prawdziwy model ogólnych prawidłowości obserwacyjnych związanych z tym zjawiskiem. Mogą dopasować wszystkie dostępne dane eksperymentalne i dokonać właściwych przewidywań. Nie oznacza to jednak, że w naszych głowach faktycznie istnieje zakodowana reprezentacja reguł (lub programu), która jest przyczynowo powiązana z powstawaniem zjawiska. Zbiór reguł gramatycznych R może na przykład poprawnie opisywać pewne ograniczenia składni języka angielskiego, ale nie oznacza to, że osoby mówiące po angielsku mają w mózgu kodowanie R, które powoduje, że wytwarzają mowę zgodną z R. Zatem nawet jeśli R może prawidłowo przewidzieć zachowanie8, niekoniecznie je wyjaśnia. To rozróżnienie jest również znane w terminologii Pylyshyna jako różnica między regułami jawnymi i ukrytymi. Ukryte reguły jedynie opisują prawidłowości behawioralne, podczas gdy reguły jawne

zakodowały reprezentacje, prawdopodobnie w naszych mózgach, które odgrywają przyczynową rolę w tworzeniu prawidłowości. Zagadnienie rzeczywistości psychologicznej rodzi poważne problemy epistemologiczne. Jakie dowody uznałyby się za potwierdzające twierdzenie, że pewne zasady są wyraźnie zakodowane w naszych mózgach? Jak odróżnić różne zestawy reguł lub różne programy komputerowe, które mimo to są opisowo równoważne? Jakim częściom skomputeryzowanego modelu należy przypisać znaczenie psychologiczne, a jakie pominąć? Ci, którzy sympatyzują z argumentami Quine'a na temat radykalnej nieokreśloności dotykającej badania języka, prawdopodobnie będą żywić podobne wątpliwości co do obliczeniowego podejścia do kognitywistyki i dojdą do wniosku, że powyższe trudności są nie do pokonania. (Chociaż nie jest konieczne akceptowanie argumentów Quine'a dotyczących nieokreśloności ani jego behawioryzmu, aby dojść do takich wniosków). Chomsky postrzega takie obawy jako przejaw empirycznych uprzedzeń na temat ludzkiego umysłu oraz głęboko zakorzenionego, ale nieuzasadnionego dualizmu metodologicznego, który zakłada ostre rozróżnienie pomiędzy sferą fizyczną i mentalną. Dla niego powyższe problemy epistemologiczne to nic innego jak zwykła kwestia indukcyjnego niedookreślenia, z którą regularnie spotykają się wszystkie nauki. Obliczeniowi kognitywiści, tacy jak Newell, Pylyshyn i inni, odpowiedzieli bardziej konkretnie, rozwijając koncepcję różnych poziomów opisu systemu. Niemniej jednak nadal istnieją poważne problemy z obliczeniowym modelowaniem poznawczym i wiele osób nadal uważa, że epistemologiczne trudności stojące przed takim modelowaniem nie wynikają ze zwykłego problemu niedookreślenia spotykanego w naukach fizycznych, ale z zasadniczo innego rodzaju problemu, który jest znacznie trudniejszy. Druga wpływowa krytyka była skierowana szczególnie przeciwko funkcjonalizmowi maszynowemu. Zostało ono przedstawione przez Blocka (1978) w formie eksperymentu myślowego, który każe nam wyobrazić sobie całą populację Chin symulujące ludzki umysł przez godzinę. Wszyscy obywatele Chin są wyposażeni w radiotelefony, które łączą ich ze sobą we właściwy sposób. Możemy myśleć o poszczególnych obywatelach Chin jak o neuronach lub jakichkolwiek elementach mózgu, które uważamy za atomowe. Ludzie są również połączeni drogą radiową ze sztucznym ciałem, z którego mogą odbierać bodźce zmysłowe i do którego mogą dostarczać sygnały wyjściowe w celu wygenerowania zachowań fizycznych, takich jak podniesienie ręki. Zgodnie z funkcjonalizmem maszynowym należałoby stwierdzić, że gdyby Chińczycy wiernie symulowali właściwą tabelę przejściową, to dzięki odpowiedniemu powiązaniu między sobą oraz z wejściami i wyjściami w rzeczywistości byliby świadomi. Wydaje nam się to jednak sprzeczne z intuicją, jeśli nie jawnie absurdalne. Powstały system mógłby być na pewnym poziomie opisu izomorficzny w stosunku do mózgu, ale nie wydaje się, aby żywił w nim jakiekolwiek odczucia, bóle, swędzenie, przekonania i pragnienia. Z podobnych powodów wynikałoby, że o żadnym czysto obliczeniowym systemie sztucznej inteligencji nie można powiedzieć, że posiada prawdziwy umysł. Niektórzy funkcjonaliści postanowili nie dać się nabrać i przyznali, że „chiński mózg" (lub odpowiednio zaprogramowany robot) w rzeczywistości posiadałby autentyczne treści umysłowe, przypisując nasze sprzeczne intuicje z szowinizmem mózgowym oraz naszą skłonność do uważania, że jedynie neurologiczne oprogramowanie mokre jest zdolne do podtrzymywanie życia psychicznego. Trudno to jednak przełknąć, a eksperyment myślowy przekonał wielu, że niepohamowany funkcjonalizm jest zbyt liberalny i należy go albo porzucić, albo znacznie ograniczyć. Trzeci przełomowy atak filozoficzny na silną sztuczną inteligencję został zapoczątkowany przez Searle'a (1980) za pomocą słynnego obecnie argumentu dotyczącego chińskiego pokoju (CRA). Agencja ratingowa wywołała ogromną ilość dyskusji i kontrowersji, dlatego tutaj przedstawimy jedynie pobieżny przegląd tej kwestii; szczegółowe omówienie czytelnika odsyłamy do Cole'a. CRA opiera się na eksperymencie myślowym, w którym występuje sam Searle. Jest w pokoju; na zewnątrz sali są rodzimi użytkownicy języka chińskiego, którzy nie wiedzą, że Searle jest w środku. Searle-in-the-room, podobnie jak Searle w prawdziwym życiu, nie zna chińskiego, ale biegle włada językiem angielskim. Osoby mówiące po chińsku wysyłają karty do pokoju przez szczelinę; na tych kartach są zapisane pytania w języku chińskim. Pokój, dzięki

uprzejmości tajnej pracy Searle'a, zwraca karty jako dane wyjściowe rodzimym użytkownikom języka chińskiego. Wyniki Searle'a powstają na podstawie zbioru przepisów: ta książka jest tabelą przeglądową, która mówi mu, jakich Chińczyków ma wyprodukować na podstawie tego, co zostało przesłane. Dla Searle'a wszyscy Chińczycy to tylko zbiór – używając języka Searle'a – zawijasów. Istota argumentu jest dość prosta: Searle-in-the-room ma być wszystkim, czym może być komputer, a ponieważ nie rozumie chińskiego, żaden komputer nie będzie w stanie tego zrozumieć. Searle bezmyślnie przesuwa squiggle-squoggles i, zgodnie z argumentacją, w zasadzie to wszystko, co robią komputery. Searle podał różne, bardziej ogólne formy argumentacji. Na przykład podsumowuje argument jako taki, w którym wychodzi się z przesłanek

1. Składnia nie wystarcza do semantyki.

2. Programy komputerowe są całkowicie określone przez ich strukturę formalną, czyli syntaktyczną.

3. Umysły mają treść mentalną; w szczególności mają zawartość semantyki , co powinno za tym podążać

Żaden program komputerowy sam w sobie nie jest wystarczający, aby nadać systemowi umysł.

Krótko mówiąc, programy nie są umysłami i same w sobie nie wystarczają do posiadania umysłów. (Na podstawie Searle 1984).

Udzielono CRA wielu odpowiedzi, zarówno w jej pierwotnym wcieleniu, jak iw ogólnej formie wyrażonej powyżej; być może dwa najpopularniejsze to odpowiedź systemu i odpowiedź robota. To pierwsze opiera się na twierdzeniu, że chociaż Searle-in-the-room nie rozumie chińskiego, cały system obejmujący go jako właściwą część tak. Oznacza to, że założenie, że Searle-in-the-room jest wszystkim, czym może być komputer, zostaje podważone. Ten ostatni zarzut opiera się na twierdzeniu, że chociaż Searle-in-the-room nie rozumie języka chińskiego, to brak ten wynika z faktu, że Searle nie jest we właściwy sposób powiązany przyczynowo ze środowiskiem zewnętrznym. Twierdzono, że w prawdziwym robocie znaczenie byłoby budowane na podstawie przyczynowych transakcji robota ze światem rzeczywistym. Zatem chociaż Searle może w pewnym sensie funkcjonować w tym pomieszczeniu jako komputer, nie funkcjonuje jako pełnoprawny robot, a silna sztuczna inteligencja ma na celu budowanie ludzi jako pełnoprawnych robotów. Searle udzielił odpowiedzi na te odpowiedzi, a kontrowersje trwają. Niezależnie od opinii na temat agencji ratingowych, argument ten niewątpliwie wywarł ogromny wpływ na tę dziedzinę. W tym samym czasie, gdy pojawiały się krytyczne uwagi filozoficzne, takie jak powyższa, zaczęły pojawiać się poważne problemy techniczne z klasyczną sztuczną inteligencją. Jednym z nich był problem z ramą. Do tej pory termin ten stał się dość niejasny. Czasami rozumiany jest jako wspomniany wcześniej problem trafności (jak rozpoznać, czy dana informacja może być istotna w danej sytuacji); czasami rozumie się, że oznacza to oczywistą niewykonalność obliczeniową holistycznych procesów myślowych; czasami jest nawet błędnie rozumiany jako ogólna etykieta niewykonalności symbolicznej sztucznej inteligencji. Być może najszersze i najmniej niedokładne jego odczytanie jest następujące: jest to problem określenia warunków, w jakich przekonanie powinno zostać zaktualizowane po podjęciu działania. W swoim pierwotnym wcieleniu problem był bardziej techniczny i wąski, pojawiał się w kontekście konkretnego zadania w określonych ramach: rozumowania o działaniu w rachunku sytuacji. Ten ostatni jest systemem formalnym, opartym na logice pierwszego rzędu, służącym do reprezentowania i wnioskowania na temat działania, czasu i zmiany. Jego podstawowym pojęciem jest płynność, czyli właściwość, której wartość może zmieniać się w czasie, np. temperatura pomieszczenia lub położenie poruszającego się obiektu. Płynni ludzie są urzeczowieni i dlatego można je określić ilościowo. Co ważne, właściwości logiczne świata same w sobie są traktowane jako płynne. Taki płynny zdaniowy może reprezentować, czy obiekt znajduje się na lewo od innego obiektu, czy też nie, lub czy światło w

pokoju jest włączone. Świat w dowolnym momencie można wyczerpująco opisać za pomocą zestawu formuł określających wartości wszystkich płynnych w tym momencie; mówi się, że taki opis reprezentuje stan świata w danym momencie. Reifikują się także działania. Każde działanie ma zestaw warunków wstępnych i skutków, z których oba są opisane w sposób płynny. Jeżeli w danym stanie spełnione są przesłanki działania, wówczas działanie może zostać przeprowadzone i skutkować będzie powstaniem nowego stanu spełniającego skutki działania. Zaczynając od stanu początkowego, który prawdopodobnie reprezentuje świat, w którym robot wchodzi do niego po raz pierwszy, możliwych jest wiele różnych sekwencji stanów, w zależności od różnych kierunków działań, które można podjąć. Aby wykluczyć dziwaczne modele, w których na przykład akcja ma skutki niezwiązane z nią, musimy wyraźnie określić brak skutków każdej akcji za pomocą tak zwanych „aksjomatów ramowych". Chociaż opracowano zwięzłe sposoby formułowania aksjomatów ramowych, wyzwaniem pozostaje złożoność obliczeniowa rozumowania na ich podstawie. Zaproponowano kilka innych proponowanych rozwiązań, począwszy od okręgu po zupełnie inne formalizmy przedstawiania i rozumowania na temat działania i zmiany. Warto zauważyć, że żadne z zaproponowanych dotychczas rozwiązań nie jest nawet w stanie zbliżyć się do efektywności, z jaką małe dzieci myślą o działaniu. Sugerowano, że ludzie nie napotykają problemu z rozumowaniem na temat braku skutków działań, ponieważ przyjmują za pewnik, że działanie nie wpływa na nic, chyba że mają dowody temu zaprzeczające. Jednak prawdziwy problem, na którym skupiali się filozofowie tacy jak Fodor, jest następujący: w jaki sposób możemy stwierdzić, czy informacja stanowi „dowód przeciwny"? Mamy tu co najmniej dwie odrębne kwestie. Najpierw musimy być w stanie określić, czy dana informacja jest potencjalnie istotna dla niektórych naszych przekonań. To jest znowu problem trafności. Po drugie, musimy być w stanie określić, czy informacja fałszuje przekonanie. Są to zarówno problemy inżynieryjne dla GOFAI, jak i ogólne problemy filozoficzne. Z punktu widzenia inżynierii zbudowanie systemu symbolicznego, który wydaje rozsądny werdykt po zidentyfikowaniu właściwych przekonań towarzyszących, nie jest zbyt trudne. Główną trudnością praktyczną jest szybkie skupienie się na istotnych informacjach. Wielu uwierzyło, że jest wysoce nieprawdopodobne, aby jakikolwiek system manipulacji symbolami mógł pokonać tę trudność.

Koneksjonizm i systemy dynamiczne

Problemy koncepcyjne i inżynieryjne, takie jak powyższe, w połączeniu z rozczarowaniem, które nastąpiło po krótkim okresie ekscytacji systemami ekspertowymi i wielkim projektem „piątej generacji" rozpoczętym w Japonii w latach 80. XX wieku, pomogły utorować drogę ostremu sprzeciwowi wobec podejścia GOFAI , zarówno w sztucznej inteligencji, jak i kognitywistyce. W dużej mierze ta reakcja przejawiła się w bardzo szybkim wzroście koneksjonizmu w latach 80. Koneksjonizm istniał co najmniej od lat czterdziestych XX wieku (podstawy założyli McCulloch i Pitts (1943)), ale dopiero w latach osiemdziesiątych zaczął wyłaniać się jako poważna alternatywa dla GOFAI, głównie dzięki wysiłkom Rumelharta, McClellanda oraz Grupa Badawcza PDP (1986). Podstawowym narzędziem koncepcyjnym i inżynieryjnym koneksjonistów jest sieć neuronowa. Sieć neuronowa składa się z szeregu węzłów (lub „jednostek") przypominających neurony mózgowe. Każdy węzeł odbiera pewną liczbę sygnałów wejściowych i dostarcza sygnał wyjściowy. Węzły są ze sobą połączone w taki sposób, że wyjście jednego węzła staje się wejściem innego węzła. Wartości wejściowe i wyjściowe są zazwyczaj reprezentowane przez liczby rzeczywiste. Połączenia mają przypisane wagi, które są również reprezentowane przez liczby rzeczywiste. Intuicyjnie waga połączenia reprezentuje wpływ, jaki jeden węzeł ma na moc wyjściową drugiego. Wynik każdego węzła jest prostą funkcją liniową wejść; zazwyczaj obliczana jest ważona suma wartości wejściowych i generowany jest wynik 1 lub 0 w zależności od tego, czy suma przekracza określony próg. Jeśli wyjście wynosi 1, mówi się, że węzeł jest aktywowany lub uruchamia się; w przeciwnym razie jest hamowany. Niektóre jednostki wyznaczane są jako węzły wejściowe i wyjściowe całej sieci; zazwyczaj istnieje tylko jeden węzeł wyjściowy. Sieci neuronowe są zdolne do pewnego rodzaju uczenia się; można je nauczyć obliczania –

lub przybliżania – funkcji celu. Istnieją algorytmy uczenia się ogólnego przeznaczenia, takie jak propagacja wsteczna, które rozpoczynając od losowych wag, wielokrotnie wystawiają sieć na różne dane wejściowe w zestawie szkoleniowym i dostosowują wagi tak, aby przybliżyć wynik do prawidłowej wartości. Skonstruowano sieci neuronowe, które dobrze radzą sobie z różnymi nietrywialnymi zadaniami poznawczymi, takimi jak nauka czasu przeszłego angielskich czasowników lub synteza mowy z tekstu pisanego. Sieci neuronowe mają wiele niezwykłych cech, które odróżniają je od systemów GOFAI. Jednym z nich jest brak centralnej jednostki przetwarzającej lub jakichkolwiek wyraźnie zakodowanych instrukcji, które określają zachowanie systemu. Istnieją tylko pojedyncze węzły, a pojedynczy węzeł ma tylko niewielką ilość całkowicie lokalnych informacji: wartości wejściowe, które otrzymuje od swoich sąsiadów. Dzięki tej ogromnej lokalizacji i wzajemnym powiązaniom sieci neuronowe są zdolne do łagodnej degradacji, co oznacza, że jeśli niektóre części sieci ulegną uszkodzeniu, sieć jako całość będzie nadal działać, a spadek wydajności będzie mniej więcej proporcjonalny do ilości szkoda. W przeciwieństwie do tego systemy manipulujące symbolami są zwykle kruche; niewielkie odchylenie od zaprogramowanego przebiegu zdarzeń może doprowadzić do katastrofalnej awarii. Taka kruchość jest nietypowa dla ludzkiej inteligencji. Podobnie jak działanie sieci neuronowych, ludzkie poznanie będzie podlegać ciągłej i łagodnej degradacji w niesprzyjających warunkach, zamiast nagłej ogólnej awarii. Po drugie, reprezentacja jest rozproszona, w tym sensie, że informacje nie są kodowane przez konkretne struktury symboliczne; raczej informacja jest zasadniczo reprezentowana jako wzór działania w całej sieci: odpalanie różnych węzłów. Ogólna „wiedza" zakodowana przez sieć neuronową zasadniczo opiera się na wagach różnych połączeń; jest subsymboliczny i wysoce rozpowszechniony. Ważnym następstwem reprezentacji rozproszonej jest to, że sieci neuronowe unikają irytującej kwestii treści, która pojawia się w przypadku klasycznego CTM. Nie pojawia się pytanie, w jaki sposób symbole atomowe nabywają swoje znaczenie – ponieważ symboli atomowych nie ma. Te interesujące cechy sieci neuronowych, w połączeniu z faktem, że wydają się one bardziej prawdopodobne biologicznie niż komputery cyfrowe, w dalszym ciągu przemawiają do wielu badaczy kognitywistyki i inżynierów sztucznej inteligencji, a intensywne badania w tej dziedzinie nie słabną, chociaż jak dotąd było stosunkowo niewiele wybitnych osiągnięć. Problem zdrowego rozsądku pojawia się jednak ponownie przy konfigurowaniu sieci neuronowych w innej formie. Inteligencja wykazywana przez (nadzorowaną) sieć neuronową jest wstępnie wbudowana w system przez człowieka, który zajmuje się modelowaniem, który szkoli sieć. Jednak to nie wystarczy, aby wystarczająco zawęzić przestrzeń możliwych hipotez, aby wykluczyć uogólnienia, które są uzasadnione z punktu widzenia danych uczących, ale nieudolne i niewłaściwe z ludzkiego punktu widzenia. Istnieją całe rzesze opowieści o sieciach neuronowych, które po intensywnym szkoleniu doszły do uogólnień, które nauczyły się rozróżniać cechy zupełnie nieistotne dla człowieka-modelującego (w rzeczywistości takie, których modelarz nawet nie zauważył). Co więcej, jeśli chodzi o moc obliczeniową, wszystko, co można zrobić za pomocą sieci neuronowych, może być wykonane przez maszyny Turinga, a zatem, zgodnie z tezą Churcha, nie ma nic, co sieci neuronowe mogą zrobić, czego nie dałoby się zrobić, powiedzmy, za pomocą programów LISP. Oznacza to, że nawet gdyby mózgi okazały się swego rodzaju gigantycznymi sieciami neuronowymi, w zasadzie możliwa byłaby symulacja jej z doskonałą precyzją przy użyciu klasycznych technik GOFAI. Wynikałoby z tego na przykład, że istnieją systemy oparte na regułach, które są w stanie przejść test Turinga, nawet jeśli systemy te są tak niewiarygodnie rozległe i nieporęczne, że praktycznie niemożliwe jest ich zbudowanie. (Chociaż gdyby mózgi okazały się niczym innym jak sieciami neuronowymi, z pewnością udowodniłoby to, że nie jest konieczne, aby system wdrażał symbolicznie zakodowaną teorię domeny opartą na regułach, aby osiągnąć kompetencje w tej dziedzinie). Istnieją inne kwestie związane z pytaniem, czy sieciom neuronowym uda się kiedykolwiek osiągnąć inteligencję ogólną, w tym słynna debata na temat systematyczności, która została zapoczątkowana przez Fodora i Pylyshyna (1988) i nadal trwa, ale nie będziemy się nimi tutaj zajmować. Ściśle powiązane z koneksjonizmem jest podejście do inteligencji

oparte na systemach dynamicznych (Port i van Gelder 1995). Podejście to czerpie z ogólnej teorii nieliniowych układów dynamicznych, traktując umysł jako ciągły układ dynamiczny – zasadniczo zbiór zmiennych, których wartości ewoluują jednocześnie w czasie. Ewolucję systemu opisuje się zwykle za pomocą zbioru praw, zwykle wyrażanych za pomocą równań różniczkowych lub różnicowych. Stan układu w danym momencie opisywany jest wartościami zmiennych w tym momencie. Wartości zmiennych w kolejnych chwilach (tj. dynamiczna trajektoria układu w przestrzeni wszystkich możliwych stanów) są zdeterminowane stanem obecnym i prawami dynamiki. Jednak teoria systemów dynamicznych jest wykorzystywana w czystej kognitywistyce, a nie w sztucznej inteligencji. Oznacza to, że zapewnia zestaw zasobów koncepcyjnych do zrozumienia poznania i modelowania jego aspektów jako systemów dynamicznych, ale nie do budowania inteligentnych systemów. W związku z tym nie będziemy mieli tu wiele do powiedzenia na ten temat (choć patrz rozdział 6). Niemniej jednak zwolennicy podejścia do poznania opartego na systemach dynamicznych zazwyczaj podkreślają znaczenie czasu, kontekstu, interakcji, ucieleśnienia i środowiska, w związku z czym są naturalnymi sojusznikami usytuowanej i osadzonej sztucznej inteligencji, do której przejdziemy dalej.

AI od dołu: zlokalizowana inteligencja

Gdy w latach 80. XX wieku zaczęło narastać rozczarowanie GOFAI, badacze sztucznej inteligencji, tacy jak Rod Brooks z MIT, doszli do wniosku, że systemy opierające się na szczegółowych symbolicznych reprezentacjach świata, przygotowanych z wyprzedzeniem przez inżynierów, oraz na generowaniu działań poprzez szczegółowe planowanie logiczne były niewykonalne, kruche i poznawczo nieprawdopodobne. Zachęcali do przeniesienia uwagi z zadań symbolicznych wyższego rzędu, takich jak rozumowanie dedukcyjne, na pozornie „proste" zadania percepcyjne i motoryczne niższego poziomu, takie jak wyczuwanie, poruszanie się, obracanie, chwytanie, unikanie przeszkód i tak dalej. Utrzymywali, że tylko w pełni ucieleśnieni agenci, zdolni do sprawnego wykonywania tych zadań, mogą zostać naprawdę uznani za sztucznych agentów i że tylko pełne wcielenie ma jakąkolwiek nadzieję na właściwe „uziemienie" sztucznego agenta w realnym świecie. GOFAI albo całkowicie zignorował, albo zminimalizował znaczenie takich działań. Zdolności percepcyjne i motoryczne były postrzegane jako zwykłe „przetworniki", peryferyjnie przydatne i istotne tylko w tym sensie, że dostarczały symboliczne reprezentacje świata do centralnych procesów myślowych lub wykorzystywały efektory do przekładania wyników takich procesów na ruchy ciała. Z pewnością tym, co odróżnia ludzi od owadów, jak twierdzi GOFAI, jest nasza zdolność do racjonalnego myślenia. Brooks i jego współpracownicy argumentowali, że zdolności cielesne nie są trywialne – w rzeczywistości GOFAI okazał się nieudolny w budowaniu systemów je posiadających. Co więcej, utrzymywali, że badanie takich zdolności może dostarczyć nam cennych spostrzeżeń na temat tego, w jaki sposób może wyłonić się z nich poznanie wyższego rzędu. Język i zdolność myślenia symbolicznego pojawiły się bardzo niedawno w historii ludzkości, co sugeruje, że ewolucja włożyła większość swojego wysiłku w budowanie naszych systemów sensorycznych i motorycznych. Kiedy zrozumiemy pozornie proste i przyziemne działanie takich systemów, zagadka inteligencji może zacząć się rozwiązywać. Język i rozumowanie staną się proste, gdy dowiemy się, jak zbudować robota, który będzie w stanie skutecznie poruszać się po świecie fizycznym, ponieważ według Brooksa „głównego składnika intelektu robota" nie można znaleźć w rozumowaniu, ale raczej w „dynamice" interakcji robota i jego otoczenia." Zasadniczo program badawczy AI realizowany przez Brooksa i jego zwolenników, który stał się znany jako sztuczna inteligencja zlokalizowana9, sprowadzał się do „patrzenia na prostsze zwierzęta jako oddolny model budowania inteligencji" . Zapożyczając sformułowanie od historyków, była to w pewnym sensie sztuczna inteligencja „od dołu". Kluczową tezą Brooksa i jego zespołu było to, że skomplikowane symboliczne reprezentacje świata są niepotrzebne do rozwiązywania szerokiego zakresu problemów. Wiele problemów można skuteczniej rozwiązać poprzez rezygnację z reprezentacji i wykorzystanie struktury otaczającego środowiska, co ujęte jest w haśle „świat jest swoją najlepszą reprezentacją".

Uważano, że ciągłe odczuwanie świata i wchodzenie w interakcję ze światem w zamkniętej pętli sprzężenia zwrotnego jest znacznie bardziej obiecującym podejściem niż budowanie jego statycznego, symbolicznego „modelu" (opisywanego, powiedzmy, jako stan w rachunku sytuacji) i rozumowanie na ten temat. Brooks i jego zespół zademonstrowali swoje podejście, budując robota o imieniu Herbert, którego zadaniem było poruszanie się po korytarzach laboratorium MIT AI Lab w celu identyfikacji i utylizacji pustych puszek po napojach. Herbert został zbudowany w oparciu o tzw. architekturę subsumpcyjną, składającą się z szeregu niezależnych modułów, z których każdy specjalizował się w wykonywaniu określonego zadania, np. posuwaniu się do przodu. W dowolnym momencie moduł może zostać aktywowany lub stłumiony, w zależności od bodźców dynamicznie odbieranych przez Herberta. Ogólny system opierał się na niewielkich lub żadnych wewnętrznych reprezentacjach i manipulacji symbolicznej, ale zdołał wykazać się zaskakująco solidnym zachowaniem. Odwracając uwagę od wewnętrznych reprezentacji i procesów w stronę zachowań zewnętrznych i ciągłej interakcji ze środowiskiem, prace Brooksa i innych badaczy sztucznej inteligencji usytuowanej oznaczają odwrotne odejście od rewolucji poznawczej i powrót w stronę behawioryzmu. Rzeczywiście, niektórzy mówili o „kontrrewolucji" w sztucznej inteligencji i kognitywistyce. Uważamy, że takie twierdzenia są przesadzone; większość badaczy tych dziedzin nie jest skłonna wyrzec się naukowej legitymizacji reprezentacji w wyjaśnianiu umysłu ani ich przydatności jako narzędzi inżynierskich. Naszym zdaniem też nie powinni. Uwagi teoretyków sytuacji zostały dobrze przyjęte i sztuczna inteligencja jako całość zwraca obecnie znacznie większą uwagę na kontekst środowiskowy i ucieleśnienie. Jest to pozytywna zmiana i tendencja ta prawdopodobnie się utrzyma. Ale istnienie reprezentacji mentalnych wydaje się teraz niezaprzeczalne, jak nigdy dotąd. Ludzie mogą po prostu zamknąć oczy, zamknąć uszy i podjąć bardzo nietrywialne myślenie o możliwych skutkach swoich działań, dziesiątej liczbie pierwszej, logicznych konsekwencjach swoich przekonań, strukturze naszego Układu Słonecznego lub cząsteczkach wody, jednorożcach i tak dalej NA. Założenie, że takie „klasyczne" myślenie stanie się proste, gdy zrozumiemy, jak wiążemy sznurowadła, jest wątpliwe i zostało zakwestionowane. Podsumowując, tabela przedstawia przybliżone graficzne przedstawienie głównych różnic między GOFAI a zlokalizowaną/ucieleśnioną sztuczną inteligencją.

Klasyczna: Ucieleśniona

reprezentacyjny: niereprezentacyjny

indywidualistyczny: społeczny

streszczenie: beton

niezależny od kontekstu: zależny od kontekstu

statyczny: dynamiczny

atomistyczny: holistyczny

inspirowane komputerem: inspirowane biologią

zorientowany na myślenie: zorientowany na działanie

Warto zauważyć, że pojawienie się teorii sytuacyjnych w sztucznej inteligencji i kognitywistyce znalazło już odzwierciedlenie – lub od tego czasu zostało odzwierciedlone – w podobnych ruchach w kilku obszarach filozofii. Na przykład w filozofii języka zwrot w stronę języka potocznego, który zaczął następować w Oksfordzie w latach pięćdziesiątych XX wieku, przede wszystkim jako reakcja na pozytywizm logiczny, można postrzegać jako prekursor behawioralnego sprzeciwu wobec kognitywizmu. W szczególności teoria aktów mowy, zapoczątkowana przez Austina, a następnie

podjęta przez Strawsona, Searle'a i innych, jako pierwsza podkreśliła, że kluczową jednostką znaczenia językowego nie jest zdanie abstrakcyjne, ale raczej wypowiedź, odwracając w ten sposób uwagę od izolowaną całość teoretyczną (zdanie) od konkretnego czynu dokonanego przez realnych ludzi w czasie rzeczywistym. Tendencja ta utrzymuje się i wzmacnia, szczególnie co widać na wykresie późniejszego rozwoju pragmatyki i rosnące uznanie rozległej i bardzo złożonej zależności znaczenia od czynników deiktycznych i innych czynników kontekstowych. Mniej więcej podobny rozwój nastąpił w filozofii nauki i matematyce, również w dużej mierze jako reakcja na pozytywizm. W nauce filozofowie tacy jak Thomas Kuhn i Joseph Agassi podkreślali, że abstrakcyjne systemy uzasadniania (zasadniczo logika indukcyjna i dedukcyjna) są kiepskimi modelami praktyki naukowej, która jest przede wszystkim działalnością ludzką w dużym stopniu uzależnioną od interakcji społecznych oraz kulturowych i czynniki polityczne. W matematyce filozofowie tacy jak Lakatos (1976) i konstruktywiści społeczni, tacy jak Barnes i Bloor (1982), rozpoczęli energiczne ataki przeciwko „euklidesizmowi", formalnemu stylowi uprawiania matematyki, w którym pozornie niewątpliwy zbiór aksjomatów stanowi podstawę i dedukcję konsekwencje są następnie wyprowadzane kumulatywnie i monotonicznie. Twierdzono, że ten styl jest zbyt schludny, aby odzwierciedlać „prawdziwą" matematykę. W rzeczywistości całkowicie pominął proces uprawiania matematyki (właściwie uzyskiwanie wyników), dynamiczne i żywe aspekty tej dziedziny. Nie twierdzimy, że wszystkie te wątki koniecznie wpływały na siebie nawzajem lub że było w którymkolwiek z nich coś nieubłaganego. Niemniej jednak niezaprzeczalne jest istnienie pewnych istotnych wspólnych punktów odniesienia i leżących u ich podstaw podobieństw. Istnieje ogólna tendencja odchodzenia od statyki w stronę dynamiki, od abstrakcji i wyrwania z kontekstu do konkretu i kontekstu, od uzasadnienia do odkrycia, od izolowanej kontemplacji do interakcji społecznych i od myślenia do działania. Dominującym i powracającym tematem było przekonanie, że prawdziwego zrozumienia nigdy nie da się osiągnąć poprzez przyjęcie czegoś dynamicznego i ewoluującego, reaktywnego, plastycznego, elastycznego, nieformalnego, zawierającego wiele niuansów, teksturowanego, kolorowego i otwartego; i modelowanie go za pomocą czegoś statycznego, rygorystycznego, nieugiętego i nieelastycznego – to znaczy zasadniczo poprzez zastąpienie czegoś żywego czymś, co martwe. Trendy, o których mowa, nie pozostają bez związku ze wzrostem znaczenia nauk społecznych, antropologii kulturowej, studiów feministycznych itd., a w dużej mierze konflikt między GOFAI a zlokalizowaną sztuczną inteligencją można postrzegać jako odzwierciedlenie niesławnych wojen naukowych, zderzenie tradycyjnej „obiektywistycznej" metafizyki z konstruktywizmem społecznym, a ostatnio, w przypadku kognitywistyki, wojny o racjonalność, podczas których teoretycy sytuacyjni kwestionują „ekologiczną ważność" klasycznych eksperymentów z zakresu kognitywistyki i psychologii oraz wzywają do większy nacisk na etologię, ważność ekologiczną i „prawdziwe" zachowanie przejawiane w „prawdziwym" świecie (w przeciwieństwie do rzekomo sztucznych i wysoce ograniczonych warunków panujących w laboratorium). Nasze uwagi tutaj nie mają na celu komentarza do wojen naukowych, ale jedynie próbę zapewnienia szerszego kontekstu dla zrozumienia ostrej reakcji na GOFAI oraz pojawienia się sytuacyjnego podejścia do sztucznej inteligencji i badania umysłu.

Przyszłość sztucznej inteligencji

Jeśli przepowiednie z przeszłości mogą stanowić jakąś wskazówkę, jedyną rzeczą, jaką dzisiaj wiemy o przyszłej nauce i technologii, jest to, że będzie ona radykalnie inna od tego, co przewidujemy. Prawdopodobnie w przypadku sztucznej inteligencji możemy już dziś wiedzieć, że postęp będzie znacznie wolniejszy, niż większość się spodziewa. Przecież na konferencji inauguracyjnej w Dartmouth College w 1956 roku Herb Simon przewidział, że myślące maszyny zdolne dorównać ludzkiemu umysłowi są „tuż za rogiem" (odpowiednie cytaty i pouczające dyskusje można znaleźć w pierwszym rozdziale Russella i Norviga 2003). Jak się okazało, nowy wiek nadejdzie bez ani jednej maszyny zdolnej do rozmowy nawet na poziomie malucha. Jeśli chodzi o budowanie maszyn zdolnych do wykazania się

inteligencją na poziomie ludzkim, lepszym prorokiem wydaje się dziś Kartezjusz, a nie Turing. Nie powstrzymuje to ludzi od dalszego publikowania niezwykle optymistycznych przewidywań. Na przykład Moravec (1999) stwierdził, że ponieważ prędkość sprzętu komputerowego podwaja się co osiemnaście miesięcy (zgodnie z prawem Moore'a, które najwyraźniej obowiązywało w przeszłości), roboty „czwartej generacji" wkrótce przewyższą ludzi pod każdym względem od prowadzenia firm po pisanie powieści. Te roboty, jak głosi historia, osiągną tak wzniosły poziom poznawczy, że staniemy wobec nich tak, jak dzisiaj stoją przed nami organizmy jednokomórkowe. Moravec nie jest bynajmniej Pollyanczykiem. Wiele innych osób zajmujących się sztuczną inteligencją przewiduje, że ta sama sensacyjna przyszłość rozwija się równie szybko. W rzeczywistości podczas obchodów pięćdziesiątej rocznicy konferencji poświęconej sztucznej inteligencji w Dartmouth w 1956 r. na uniwersytecie gospodarz i filozof Jim Moor postawił pytanie: „Czy sztuczna inteligencja na poziomie ludzkim zostanie osiągnięta w ciągu najbliższych 50 lat?" do pięciu myślicieli, którzy wzięli udział w pierwotnej konferencji: Johna McCarthy'ego, Marvina Minsky'ego, Olivera Selfridge'a, Raya Solomonoffa i Trencharda Moore'a. McCarthy i Minsky stanowczo i bez wahania potwierdzili, a Solomonoff zdawał się sugerować, że sztuczna inteligencja zapewniła jedyny promyk nadziei w obliczu faktu, że nasz gatunek wydaje się pragnąć samozniszczenia. (Odpowiedź Selfridge'a była nieco zagadkowa. Moore odpowiedział stanowczo, jednoznacznie przecząco i oświadczył, że kiedy jego komputer będzie na tyle inteligentny, aby móc z nim rozmawiać na temat problemów matematycznych, będzie mógł potraktować całe to przedsięwzięcie poważniej). Pytanie Moora nie jest przeznaczone tylko dla niego. naukowcy i inżynierowie; jest to także pytanie do filozofów. Dzieje się tak z dwóch powodów: (1) badania i rozwój mające na celu potwierdzenie odpowiedzi twierdzącej muszą obejmować filozofię z powodów podanych powyżej; (2) filozofowie mogliby teraz przedstawić argumenty, które ostatecznie odpowiedzą na pytanie Moora. Jeśli którakolwiek z ostrych krytyk AI, które omówiliśmy, jest zasadniczo słuszna, to oczywiście AI nie będzie w stanie wyprodukować maszyn posiadających zdolności umysłowe ludzi. W każdym razie czas płynie dalej i pokaże.