Wyzwania filozoficzne

Kartezjusz utrzymywał, że naszym rozumem jest „powszechny instrument." Ponieważ wierzył, że każdy mechanizm musi coś mieć specjalnego przeznaczenia i że żaden zbiór mechanizmów specjalnego przeznaczenia nie może być wystarczająco duży, aby objąć wszystko, co może zrobić rozum, doszedł do wniosku, że żaden mechanizm nie jest w stanie urzeczywistnić ludzkiego rozumu. Akwinata (1265–72, I, Q.75, a. 2) również argumentował, że intelektu nie zapewnia narząd materialny. Wierzył, że gorzki humor wywołany chorobą może zakłócać smak słodyczy lub każdego smaku innego niż gorzki. Analogicznie sądził, że gdyby nasze intelekty były materialne, nie mogłyby one poznawać rzeczy materialnych o różnej naturze. Większość współczesnych filozofów zaakceptowałaby fakt, że naszą inteligencję zapewniają materialne mózgi, i dlatego nie byłaby skłonna kwestionować możliwości istnienia urządzeń sztucznie inteligentnych na podstawie ich materialności. Pozostałe pytania i problemy dotyczące sztucznej inteligencji można podzielić na te, które są w dużej mierze niezależne od poszczególnych podejść do sztucznej inteligencji, oraz te, które wynikają z bardziej szczegółowych pomysłów na temat sztucznie realizowanych architektur poznawczych. Zaczniemy od kwestii bardziej ogólnych.

Ogólne pytania

Wstępy terminologiczne

Powszechnie zgodzono by się, że do wykonania niektórych zadań wymagana jest inteligencja – na przykład znalezienie ilorazu 231 podzielonego przez 42. Można powiedzieć, że urządzenie ma sztuczną inteligencję na wypadek, gdyby można było za jego pomocą wykonać zadanie, które zgodzić się, że do egzekucji przez człowieka będzie wymagana inteligencja. W tym sensie posiadanie inteligencji będę nazywał „inteligencją zadaniową". Kalkulatorów można użyć do znalezienia ilorazu 231 podzielonego przez 42, więc najwyraźniej nawet kalkulatory mają inteligencję zadaniową. Sztuczna inteligencja zadaniowa nie budzi kontrowersji; nawet Akwinata czy Kartezjusz nie mieliby powodu sprzeciwiać się inteligencji zadaniowej. Musimy jednak zauważyć, że kalkulatory nie znają odpowiedzi na problemy arytmetyczne ani nie wiedzą, że wykonują arytmetykę. Jasne jest zatem, że nie mamy prawa przechodzić od posiadania inteligencji zadaniowej do większych twierdzeń o inteligencji rzeczy, którą możemy wykorzystać do wykonania zadania. Zamierzonym przeciwieństwem inteligencji zadaniowej jest „inteligencja rzeczowa". Jeśli urządzenie ma sztuczną inteligencję, to ono – to urządzenie – jest inteligentne. Kontrowersyjne współczesne pytania dotyczą możliwości sztucznej inteligencji rzeczy i różnych podejść do jej projektowania. Możemy rozróżnić dwa gatunki inteligencji rzeczy, zapożyczając terminologię ze stacji benzynowej. Powiedzmy, że urządzenie ma sztuczną inteligencję rzeczy premium (lub „AI premium"), jeśli ma swoją inteligencję w sposób, który wyjaśnia, w jaki sposób ludzie mają inteligencję rzeczy. Jeśli urządzenie posiada inteligencję w sposób, który nie wyjaśnia, w jaki sposób inteligencja jest u ludzi, powiedzmy, że posiada ono zwykłą, sztuczną inteligencję rzeczy (lub „zwykłą sztuczną inteligencję"). Należy zauważyć, że zwykła benzyna jest prawdziwą benzyną; podobnie zwykła sztuczna inteligencja to prawdziwa inteligencja. Zgodnie z tymi definicjami mogłoby się okazać, że nie ma realnej różnicy pomiędzy sztuczną inteligencją zwykłą a premium; to znaczy, mogłoby się okazać, że jedynym sposobem zapewnienia inteligencji rzeczy w urządzeniu jest dostarczenie jej w sposób, który pouczałby o tym, skąd posiadamy inteligencję. Ale nie jest to prawdą aprioryczną, że tak jest, dlatego potrzebujemy obu terminów. Searle (1980) wprowadził rozróżnienie pomiędzy silną i słabą sztuczną inteligencją. Silna sztuczna inteligencja to teza, że „odpowiednio zaprogramowany komputer naprawdę jest umysłem w tym sensie, że można dosłownie powiedzieć, że komputery, którym podano odpowiednie programy, rozumieją i mają inne stany poznawcze" (s. 417; podkreślenia w oryginale). Słaba sztuczna inteligencja twierdzi jedynie, że komputery stanowią użyteczne narzędzie do rygorystycznego formułowania i testowania hipotez na temat umysłu. To rozróżnienie Searle'a

pozwoliło mu łatwo skoncentrować swoją argumentację, która, jak zobaczymy później, była skierowana wyłącznie przeciwko mocnej tezie o sztucznej inteligencji. Jednak dla bardziej ogólnych celów potrzebujemy dalszych, właśnie wprowadzonych rozróżnień. Rozróżnienie regularne/premium pozwala nam rozpoznać dwie możliwości w ramach silnej sztucznej inteligencji. Rozróżnienie zadanie/rzecz pozwala nam zgodzić się z większością autorów w kwestii tego, co liczy się jako praca w badaniach nad sztuczną inteligencją, uznając jednocześnie, że wyzwania filozoficzne są zazwyczaj ukierunkowane na twierdzenia, że urządzenie samo w sobie może być rzeczywiście inteligentne lub posiadać własne stany poznawcze. Ponieważ nie ma prawdziwego problemu co do możliwości inteligencji zadaniowej, niekwalifikowane wystąpienia „inteligencji" w poniższym tekście zawsze będą oznaczać „inteligencję rzeczy".

Test Turinga

Jeśli założymy, że możliwe jest stworzenie sztucznie inteligentnego urządzenia, pojawia się pytanie, skąd moglibyśmy wiedzieć, że udało nam się je wyprodukować. Choć Alan Turing (1950) sugerował, abyśmy zastąpili takie pytania pytaniami o wydajność w jego „grze w naśladownictwo", jego propozycję często traktowano jako test, którego „zdanie" wskazywałoby, że urządzenie posiada inteligencję. W swojej standardowej interpretacji Test Turinga (jak stał się znany) składa się z serii prób, podczas których przesłuchujący wchodzi w interakcję zarówno z maszyną, jak i człowiekiem, i po pewnym określonym, skończonym czasie wydaje orzeczenie: „ X to człowiek, Y to maszyna" lub „Y to człowiek, X to maszyna". Przykłady Turinga (od analizy poezji po problemy arytmetyczne) pokazują, że zakłada się, że przesłuchujący rozumieją, że ich zadaniem jest poprawna identyfikacja i że mają mieć jak największą swobodę w zakresie pytań. Jedynym ograniczeniem, jakie nałożył na wymagane występy, było to, że można je było wystawiać za pośrednictwem telegrafu. 50-procentowy wskaźnik powodzenia deklaracji przesłuchujących oznaczałby, że nie nauczyli się oni niczego, porównując reakcje ludzi i maszyn na swoje sondy, a odsetek zbliżony do tego prawdopodobnie liczyłby się jako „zdanie" testu Turinga. Zakładając, że zgodzi się, że angażowanie się w rozmowę wymaga szerokiego wykorzystania naszej inteligencji, można wysunąć argument, że skuteczne urządzenia (urządzenia, które „przechodzą") ćwiczą inteligencję. Taka zdolność nie byłaby sklasyfikowana jako zwykła inteligencja zadaniowa, ponieważ zakres dopuszczalnych sond jest tak szeroki. Co ciekawe, według Kartezjusza artefakty nigdy nie byłyby w stanie osiągnąć zdolności konwersacyjnej. Dla kontrastu Turing (1950) dokonał słynnej prognozy, że do roku 2000 będą istnieć urządzenia, w przypadku których przesłuchujący po pięciu minutach przesłuchania nie będą w stanie uzyskać więcej niż 70% poprawności. Tę nieudaną prognozę należy oczywiście odróżnić od poglądu, że dobre wyniki w teście Turinga oznaczałyby, że urządzenie posiadało inteligencję, gdyby rzeczywiście osiągnięto taki wynik. Od 1990 r. coroczny konkurs o Nagrodę Loebnera ma na celu stymulowanie działań zmierzających do osiągnięcia tego celu. Wysunięto alternatywne interpretacje opisu procedury Turinga i wysunięto kilka sugestii dotyczących jej ulepszenia. W 2000 roku czasopismo Minds and Machines opublikowało numery z okazji pięćdziesiątej rocznicy powstania, które zawierają dyskusje na temat kilku z tych interpretacji i sugestii. Ned Block (1981) opisał maszynę, która wydaje się podważać akceptowalność testu Turinga. Pracownicy zaczynają od wyszczególnienia wszystkich zagrań, od których może rozpocząć przesłuchujący. Otwieracze te są ułożone alfabetycznie i dla każdego z nich tworzona jest odpowiedź dla maszyny w taki sposób, że kombinacja otwieracza przesłuchującego i odpowiedzi maszyny jest wiarygodna jak rozmowa ludzka. Następnie pracownicy sporządzają listę i porządkują alfabetycznie wszystkie możliwe kolejne dochodzenia prowadzone przez przesłuchujących. (Ponieważ przesłuchujący mogą zmieniać temat, kiedy tylko zechcą, późniejsze dociekania będą obejmować prawie wszystkie wcześniejsze.) Ponownie, odpowiedzi są ułożone w taki sposób, że każda połączona seria otwieracza + odpowiedź + dalsze sondowanie + odpowiedź jest wiarygodna jako rozmowa międzyludzka. Robotnicy kontynuują w ten sposób, aż do osiągnięcia określonej, skończonej długości

rozmów. Podczas testu Turinga maszyna po prostu wyszukuje sondy przesłuchującego na alfabetycznej liście i zwraca zapisane odpowiedzi. Argument wypływający z tego opisu jest taki, że maszyna Blocka przeszłaby test Turinga, ale ewidentnie nie jest inteligentna. (To tylko puszkowanie plus zwykłe urządzenie sprawdzające i nie ma większej inteligencji niż system czytnika kodów kreskowych w sklepie spożywczym.) Dlatego test Turinga nie jest dobrym testem na inteligencję. Osoby odmienne od tego argumentu mogą przyznać, że logiczna możliwość istnienia maszyny Blocka (znanej również jako Blockhead) uniemożliwia użycie testu Turinga jako definicji inteligencji. Mogą jednak nadal twierdzić, że ogrom przestrzeni możliwych sond powoduje, że prawdopodobieństwo faktycznego istnienia maszyny Blocka jest znikomo małe i że w konsekwencji procedura Turinga może nadal być wysoce skutecznym testem inteligencji.

Cel, świadomość i intencjonalność

Inteligentne zachowanie jest celowe; implikuje jakiś cel, choćby tylko udzielenie poprawnej odpowiedzi na pytanie. Trudno sobie wyobrazić, że jakieś działanie ma sens, jeśli wcześniej czy później czyjeś przyjemne lub nieprzyjemne doznania cielesne, emocje, wzruszenie religijne, uniesienie lub depresja nie są zależne od tego działania. Prawdopodobnie żaden z terminów określających takie stany nie może być w pełni zrozumiany przez istoty, które nie mogą znajdować się w żadnym z nich. Stany te są jednak stanami świadomymi. Można zatem stwierdzić, że sztuczna inteligencja wymagałaby sztucznej świadomości. Ponieważ większość podejść do sztucznej inteligencji nie ma na celu wytwarzania świadomości, można wątpić, czy typowe podejścia do sztucznej inteligencji mogą naprawdę odnieść sukces. Większość badaczy sztucznej inteligencji byłaby jednak usatysfakcjonowana, gdyby mogli zapewnić odpowiednie odpowiedzi na cele, przy czym cele są takie same jak cele, z tą różnicą, że nie wymagają świadomości. Oznacza to, że cele funkcjonują jako stany celów, w odniesieniu do których można organizować działania. Celem maszyny do gry w szachy jest wygrywanie, a sukces według tego standardu będzie prawdopodobnie liczony jako sukces dla zadania AI, nawet jeśli nikt nie przypuszcza, że maszyna odczuwa radość po wygranej lub jest opuszczona po przegranej. Gdyby zakres zadań, do których miałyby zastosowanie podobne uwagi, byłby wystarczająco szeroki i zróżnicowany, inteligencja rzeczy prawdopodobnie zostałaby osiągnięta. Na przykład robot, który mógłby robić zakupy spożywcze, w tym oddzielać dojrzałe owoce od niedojrzałych, dostosowywać się do zmian w rozkładach jazdy autobusów, opóźnień spowodowanych pogodą itp., można by prawdopodobnie uznać za sztucznie (rzecz) inteligencję, bez żadnych przypuszczenie, że ma świadome gusta lub poczucie satysfakcji z dobrze wykonanej pracy. Głębsza linia argumentacji głosi, że myśli mogą dotyczyć rzeczy tylko wtedy, gdy są produktami ewolucyjnej historii, w której przetrwanie funkcjonuje jako naturalnie przewidziany cel (Dretske 1995). Ponieważ robotom brakuje wymaganej historii ewolucji, można dojść do wniosku, że ich „myśli" nie mogą tak naprawdę dotyczyć niczego. Kontrowersyjne jest jednak to, czy taka historia jest wymagana dla prawdziwej intencjonalności (aboutness). I nawet gdyby uznano, że jest to wymagane, wielu myślicieli chętnie poprzestałoby na intencjonalności*, czyli ogłosiłoby sukces w zapewnieniu sztucznej inteligencji każdemu robotowi, który potrafiłby nie tylko dobrze mówić, ale także dopasowywać swoje działania do słów jednocześnie angażując się w rzeczywiste obiekty i procesy w czasie rzeczywistym. Posiadanie intencjonalności (lub nawet intencjonalności*) sugeruje posiadanie wewnętrznych stanów lub zdarzeń, które reprezentują doczesne przedmioty i ich właściwości. Zobaczymy, że pytania dotyczące reprezentacji i ich wykorzystania odgrywają ważną rolę w wielu wyzwaniach stojących przed sztuczną inteligencją. Dalsze wątpliwości co do intencjonalności maszyn i inteligencji robotów podniósł Searle (1992). Jego zdaniem obliczenia nie są nieodłączną częścią żadnego systemu fizycznego; zamiast tego jest „względny obserwatora", to znaczy przypisywany przez posiadaczy prawdziwej intencjonalności, takich jak my. Możemy na przykład spojrzeć na włącznik światła jak na komputer, przypisując 0 do „Off", 1 do „On" i traktując każdy ruch przełącznika jako przyjmujący 0 lub 1 jako wejście i dający odpowiednio 1 lub 0

jako wyjście . Dokonując odpowiedniego wyboru zadań, każdy wystarczająco złożony system można uznać za obliczający wszystko, co nam się podoba. Ponieważ, jak to ujął Searle, składnia nie jest nieodłącznym elementem fizyki, żaden fizyczny opis stanów robota nie mógłby określić, czy którykolwiek z nich dotyczy czegokolwiek, nawet gdybyśmy mogli wybrać zadanie, które uczyniłoby robota dla nas użytecznym. Zwolennicy poglądu Searle'a twierdzą w skrócie, że można osiągnąć niearbitralne przypisanie znaczeń językowym i pozajęzykowym zachowaniom robota, pod warunkiem, że robot będzie wystarczająco złożony i kompetentny; że uwzględniamy relacje przyczynowe między jej stanami a otoczeniem zewnętrznym; oraz że dodatkowo uwzględnimy, jakie byłoby jego zachowanie w warunkach alternatywnych. Sukces w tym zakresie wspierałby przypisywanie inteligencji rzeczy takiemu urządzeniu. Przydatne dalsze dyskusje na temat tej nieco technicznej kwestii można znaleźć u Chalmersa (1996) i Piccininiego (2010).

Mechanizm a racjonalność

Inteligencja godna tego miana powinna zbliżać się do racjonalności. Czasami uważa się, że racjonalność stoi w konflikcie z mechanizmem. Jeśli rzeczywiście taki konflikt będzie istniał, to sztuczna inteligencja nie będzie możliwa. Domniemany konflikt polega na tym, że racjonalność jest normatywna, a przyczynowość w rodzaju komputerów i robotów nie. Wyjaśnienie: Bycie racjonalnym wymaga postępowania według norm prawidłowego wnioskowania, w tym przypadków dedukcyjnych (np. akceptowanie modus ponens i odrzucanie afirmacji następstwa) i przypadków indukcyjnych (np. wnioskowanie o wysokim prawdopodobieństwie na podstawie dużej częstotliwości obserwowanych próbek lub stosowanie zasady cały materiał dowodowy). Ale przyczynowość mechaniczna jako taka nie respektuje norm logicznych; równie łatwo jest spowodować, aby urządzenie wydrukowało 2 + 2 = 5, jak spowodować, że wyświetli ono poprawne stwierdzenie. Alternatywne sformułowanie tego samego punktu idzie w ten sposób. Istotę racjonalną można scharakteryzować jako taką, która dostosowuje swoje przekonania w świetle dowodów, ponieważ dowody te są rozumiane jako istotne dla danego przekonania. Ale komputery i roboty zmieniają swoje stany wewnętrzne tylko ze względu na swoje wejścia i okablowanie. Dlatego żadnej z ich zmian stanu nie można uznać za przejaw racjonalności. W odpowiedzi na to wyzwanie zwolennicy sztucznej inteligencji mogą przyznać, że klasa sztucznych urządzeń zawiera wiele przykładów, które nie szanują racjonalności. Jednakże inni członkowie tej klasy mogą być projektowani w taki sposób, że ich operacje mechaniczne odpowiadają prawidłowym zasadom wnioskowania. (Operacje mechaniczne to zmiany lub serie zmian w układach fizycznych, które zachodzą zgodnie z prawami natury. Układy mechaniczne – np. mechanizmy wahadłowe – są paradygmatycznie mechanistyczne, ale układy elektryczne, układy neuronowe itp. są mechanistyczne, ale nie są mechaniczne) Zwolennicy sztucznej inteligencji mogą utrzymywać, że ewolucja zaprojektowała nasze mózgi właśnie w ten sposób – to znaczy, że nasze mózgi to mechanizmy neuronowe uporządkowane w taki sposób, aby ograniczać nasze przekonania zasadami logicznymi. O tym, że taka konstrukcja jest możliwa, świadczą urządzenia tak proste jak kalkulatory, które wyraźnie działają mechanicznie i jednocześnie wyraźnie respektują zasady arytmetyki. Z tego punktu widzenia problem sztucznej inteligencji polega na znalezieniu projektów, których działanie mechaniczne jest równoległe do stosowania zasad logicznych w szerokim zakresie przypadków. (W przypadku sztucznej inteligencji nie jest wymagana doskonała racjonalność, ponieważ ogólnie przyjmuje się, że istoty ludzkie są inteligentne, ale nie doskonale racjonalne). Zmiany stanu urządzeń tego rodzaju nie mogą być spowodowane relacjami logicznymi lub dowodowymi; będzie jednak utrzymywane, że równie dobrze będzie, jeśli projekt uzupełni swój zbiór przekonań tylko w przypadkach, gdy zachodzi właściwa relacja logiczna lub dowodowa. Na poparcie tej koncepcji można argumentować, że tak właśnie musi być z ludźmi. Relacje logiczne i dowodowe są bytami abstrakcyjnymi i dlatego nie mogą być przyczynami poszczególnych zdarzeń. Jedynie urządzenia, których konstrukcja ogranicza zdarzenia same w sobie, aby odpowiadały logikom i zasadom dowodowym, mogą mieć przyczyny zmian, które

respektują te zasady. Wolną wolę często kojarzono z racjonalnością (Aquinas 1265–72/1945, I, Q83, a. 1; Hasker 1999), a to skojarzenie może stanowić wyzwanie dla sztucznej inteligencji. Robotami, niezależnie od ich wyrafinowania, rządziłyby deterministyczne prawa, dlatego niektórzy mogliby powiedzieć, że brakuje im racjonalności, ponieważ zgodnie z jedną z tradycji systemom deterministycznym brakuje wolnej woli. Oczywiście takie stanowisko byłoby nie do pogodzenia z uznaniem ludzkiej inteligencji za całkowicie wyjaśnioną poprzez odniesienie do aktywności naszych materialnych mózgów. Istnieje jednak długoletnia tradycja „kompatybilistyczna", zgodnie z którą wolna wola nie jest sprzeczna z mechanizmem deterministycznym. (Hume 1748, viii; Ayer 1954; Dennett 1973). Ogólnie rzecz biorąc, tradycja ta utrzymuje, że wolna wola jest obecna, gdy procesy rozumowania wywierają swój normalny wpływ na zachowanie. Pozbawienie wolności nastąpi tylko wtedy, gdy procesy rozumowania (takie jak wyciąganie konsekwencji działań) zostaną zakłócone poprzez uszkodzenie, ominięcie lub unieważnienie brutalną siłą. Wydaje się, że to rozróżnienie między procesami rozumowania, które działają normalnie, a procesami rozumowania, w które zakłóca się na różne sposoby, ma zastosowanie do działania robotów. Zatem, jeśli tradycję kompatybilizmu uda się utrzymać na niezależnych podstawach, podejrzenie o konflikt pomiędzy inteligencją robotyczną a wolną wolą może zostać rozwiane.

Argumenty Gödla

Gödel (1931) udowodnił wynik matematyczny, który czasami uważano za sugerujący ograniczenie inteligencji maszyn. Aby zrozumieć wiarygodność tego poglądu, możemy zacząć od wyobrażenia sobie sposobu traktowania arytmetyki wzorowanego na podejściu Euklidesa do geometrii. Oznacza to, że możemy wyobrazić sobie system rozpoczynający się od pewnych aksjomatów, które (biorąc pod uwagę standardowy schemat interpretacji znaków w systemie) dotyczą dodawania i mnożenia. Na przykład jednym z aksjomatów może być wzór, którego standardowa interpretacja brzmi: „Istnieje liczba n taka, że dla dowolnej liczby m n razy m równa się m". Dalszy rozwój systemu składałby się wówczas z formuł wyprowadzonych z aksjomatów, których standardowymi interpretacjami byłyby twierdzenia arytmetyczne. Gödel zajmował się systemami formalnymi, to znaczy systemami, w których dowód można zdefiniować w kategoriach jednoznacznych zasad dodawania formuł do systemu. Przykładem takiej reguły jest to, że jeśli zdania p i p → q są już twierdzeniami układu, to q można dodać do listy twierdzeń. Intuicyjnie dwie właściwości wydają się pożądane w formalnym systemie arytmetycznym. Po pierwsze, powinno być spójne; to znaczy, że nasz system nie powinien zawierać dowodów na dwie formuły, których standardowe interpretacje są ze sobą sprzeczne. Po drugie, chcielibyśmy, żeby nasz system taki był kompletny; to znaczy, chcielibyśmy mieć system taki, że dla każdego prawdziwego wyrażenia arytmetycznego istnieje formuła, którą można wyprowadzić w systemie, której standardową interpretacją jest to stwierdzenie. Zdumiewającym odkryciem Gödla było to, że żaden formalny system arytmetyczny nie może być jednocześnie spójny i kompletny. Aby skondensować długą i trudną historię, Gödel pokazał, w jaki sposób każdemu wzorowi proponowanego systemu formalnego arytmetyki można przypisać unikalny numer („liczbę Gödla"). Następnie pokazał, że dowolny system przewidujący pewną część arytmetyki można wykorzystać do skonstruowania wzorów, których standardowa interpretacja wskazywałaby, że wzór z liczbą Gödla N nie jest twierdzeniem tego systemu. Wreszcie udało mu się wykazać, że dowolny system s, który przewiduje tę część arytmetyki, będzie zawierał pewien wzór (nazwijmy go G(s), zdanie Gödla dla systemów s), którego interpretacja będzie brzmiała: „Wzór z liczbą Gödla N nie można udowodnić w s" i którego liczbą Gödla jest N. Gdyby ten wzór można było udowodnić w s, wówczas oba formuły, które można standardowo interpretować jako „G(s) można udowodnić" i „G(s) nie można udowodnić" byłyby możliwe do wyprowadzenia, i system byłby niespójny. Jeśli jednak G(s) nie da się udowodnić w systemie, to to, co jest w nim napisane, jest prawdziwym stwierdzeniem arytmetycznym, którego nie można wyprowadzić w systemie, a zatem system jest niekompletny. Argumentacja oparta na tym

wyniku rozpoczyna się od obserwacji, że każdy mechanizm można przedstawić jako system formalny, w którym wyniki mechanizmu odpowiadają twierdzeniom tego systemu formalnego. Tak więc, ponieważ wszystkie systemy formalne muszą być albo niespójne, albo niekompletne dla arytmetyki, każdy mechanizm musi albo dawać niespójne wyniki, albo nie być w stanie wygenerować wszystkich prawd arytmetyki. W szczególności mechanizm nie byłby w stanie wygenerować G(ów), gdzie „s" jest systemem formalnym reprezentującym ten mechanizm. Jednakże niektórzy myśliciele argumentowali, że wykorzystując idee zawarte w metodzie dowodzenia twierdzenia Gödla, matematycy mogliby zrozumieć prawdziwość każdego twierdzenia arytmetycznego, w tym zdania Gödla dowolnego systemu formalnego, jaki można by zaproponować jako reprezentującego ich zdolności poznawcze. I mogliby to zrobić, nie zaprzeczając tym samym sobie. Jeśli to prawda, ludzkie zdolności poznawcze przewyższają możliwości jakiegokolwiek możliwego mechanizmu. Wielu filozofów krytykowało różne aspekty tej próby zastosowania dzieła Gödla. Kluczowy punkt sprzeciwu zaczyna się od obserwacji, że rzekome ograniczenie możliwości mechanizmów wymaga wykazania, że jest coś, co możemy zrobić, ale maszyny nie. Jednakże zarówno mechanizm reprezentowany przez systemy s, jak i my (jeśli przestudiujemy pracę Gödla) możemy udowodnić, że 1 Jeśli s jest spójne, to s nie może udowodnić G(s). Ale ani my, ani mechanizm reprezentowany przez s nie możemy udowodnić, że 2 s nie może udowodnić G(s). Oczywiście, gdyby matematycy mogli udowodnić, że są one spójne, mogliby argumentować, że zastąpienie „s" umożliwiłoby im wyprowadzanie (2). Prowadziłoby to do sprzeczności (ponieważ G(s) twierdzi, że s nie może dowieść G(s)), w związku z czym mogliby odrzucić założenie, że można je zastąpić „s". Ponieważ w tym argumencie s może być dowolnym mechanizmem, w rezultacie matematycy-ludzcy mogliby zawsze zasadnie odrzucić przypuszczenie, że są one równoważne mechanizmowi. Jednakże ta linia obrony opiera się na założeniu, że matematycy mogą udowodnić, że są konsekwentni. Nie jest wcale oczywiste, że da się to zrobić. Co więcej, innym wynikiem Gödla jest to, że jeśli system formalny jest spójny, nie może zawierać dowodu, że tak jest. Jeśli nie możemy udowodnić, że jesteśmy konsekwentni, pozostaje otwarte, że (1) można udowodnić zarówno przez maszyny, jak i przez nas, oraz (2) nie da się udowodnić przez żadne z nich; i różnica między tym, co jest możliwe dla nas, a tym, co jest możliwe dla maszyn, nie zostanie pokazana.

Klasyczne podejście do AI

Jeśli uważa się, że wątpliwości co do możliwości inteligencji maszynowej są uzasadnione, pojawia się pytanie, w jaki sposób można ją wytworzyć. W tej i dwóch kolejnych sekcjach rozważymy trzy wiodące podejścia do tego pytania. Klasyczne podejście do sztucznej inteligencji (znane również jako GOFAI, czyli dobra, staromodna sztuczna inteligencja, za Haugelandem 1981) rozwinęło się z czynników omówionych w poprzednich sekcjach. Zasady logiczne (lub reguły) mają zastosowanie do zdań i, ogólnie rzecz biorąc, zależą od terminów i wewnętrznej struktury tych zdań. (Na przykład zaakceptowanie stwierdzenia „Wszystkie nietoperze są ssakami, wszystkie ssaki są stałocieplne, zatem wszystkie nietoperze są stałocieplne" zależy od docenienia siły „wszystkich" oraz tożsamości i porządku terminów podmiotowych i orzeczeniowych występujących w Wydaje się zatem naturalne, aby pojmować inteligencję (czy to naturalną, czy sztuczną) jako obejmującą stany, które odpowiadają terminom i strukturom terminów oraz regułom działania na tych terminach i strukturach. Reguły wydają się bardzo naturalnie zawarte w programach – w istocie formę „Jeśli warunek X jest spełniony, wykonaj Y, w przeciwnym razie wykonaj Z" można opisać jako regułę, aby wykonać Y, jeśli X i Z, jeśli nie X. Kontynuując tę koncepcję, jeden potrafi traktować wpisy danych jako (ustrukturyzowane) reprezentacje faktów, a programy jako sposoby urzeczywistniania reguł manipulacji takimi reprezentacjami. Klasyczne podejście do sztucznej inteligencji napotkało wiele dobrze znanych wyzwań, które teraz rozważymy.

Pokój chiński Searle'a

Schank i Abelson (1977) powiązali rozumienie ze skryptami, które składają się z zapisanej wiedzy o strukturze konkretnych sytuacji – na przykład jedzenia w restauracji czy pójścia na przedstawienie teatralne. Krótka uwaga lub historia może przywołać scenariusz, a wiedza zawarta w scenariuszu może zostać wykorzystana do określenia oczekiwań i reakcji. Schank i Abelson utrzymywali, że nabywamy wiele pism i że „większość zrozumienia opiera się na pismach" . Prace Schanka i Abelsona stały się podstawą słynnego artykułu Johna Searle'a (1980). Searle wyobraził sobie siebie w pokoju zawierającym skrypty w języku chińskim, opowiadania w języku chińskim i program (w języku angielskim), które pozwalały na operacje na chińskich znakach wyłącznie ze względu na ich kształt (tj. nie zapewniano żadnych tłumaczeń ani środków do wykonywania tłumaczeń). Searle wyobraził sobie, że otrzymuje pytania napisane po chińsku i uruchamia program. Czasami program nakazywał skopiowanie kształtu na kartkę papieru, a seria takich kształtów była ostatecznie przekazywana rodzimym użytkownikom języka chińskiego poza pokojem. Searle przyznał, że odpowiedzi te mogą być tak dobre (z punktu widzenia pytających), jak tylko się chce, tak aby ludzie z zewnątrz mieli wszelkie powody, by sądzić, że coś na sali zrozumiało pytania. Ale tak naprawdę Searle nie znał chińskiego i nic nie rozumiał; z jego punktu widzenia nie było nic innego jak identyfikowanie zawijasów i zawijasów w różnych pozycjach i sekwencjach oraz okazjonalne kopiowanie kształtów na wydrukowany papier. Jeśli ktoś rozda gazetę z napisem po chińsku: „Krzycz, jeśli chcesz hamburgera", Searle mógłby rozdać gazetę z Chińczykami: „Tak, chcę hamburgera", ale nie miałby powodu krzyczeć, nawet jeśli był potwornie głodny. Argument Searle'a dopuszcza możliwość pewnego rodzaju osiągnięcia i wygodnie będzie mieć określenie na to osiągnięcie. W tym celu wprowadzam termin „elastyczność" w następujący sposób.

"X has flexibility" $=_{df}$ X can respond appropriately to a wide range of novel circumstances.

„Odpowiednie", „szeroki zakres" i „nowość" nie są precyzyjnymi terminami, ale nie są one oczywiście mniej precyzyjne niż „inteligencja" i wydaje się, że elastyczność, tak jak ją właśnie zdefiniowano, obejmuje przynajmniej część tego, czego można by oczekiwać od wszystkiego, o czym mówi się, że posiada inteligencję. To, czego Searle w charakterystyczny sposób odmawia programom, to nie elastyczność, ale zrozumienia i jednym ze sposobów wyrażenia swojego wniosku jest to elastyczność można osiągnąć bez zrozumienia. Innym jest to, że formalna manipulacja symbolami (tj. manipulacja symbolami wyłącznie na podstawie ich kształtów) jest nieadekwatna do zadania polegającego na zapewnieniu zrozumienia. Sama biegłość syntaktyczna, jakkolwiek właściwa i szeroka, nie może zapewnić treści semantycznej (znaczenia; zrozumienia). Wykazanie elastyczności nie jest zatem demonstracją inteligencji, jeśli przyjmuje się, że „inteligencja" wymaga zrozumienia. Formalna manipulacja symbolami jest dokładnie tym, co robią komputery, zatem to, co robią komputery, jest niewystarczające do zapewnienia zrozumienia. Mogą dostarczyć słów, które my, którzy mamy zrozumienie, możemy uznać za odnoszące się do rzeczy na świecie; ale dla nich nie ma intencjonalności – to znaczy dla nich ich symbole to tylko kształty i w ogóle nie o niczym. Searle rozważył szereg zarzutów wobec swojej argumentacji. Najważniejszym z nich jest odpowiedź systemu. Możemy myśleć o odpowiedzi systemowej jako o twierdzeniu o tym, gdzie wyznaczyć najkrótszą granicę osoby rozumiejącej; mianowicie ma być rozciągnięty wokół całego systemu, gdzie cały system obejmuje nie tylko człowieka w środku, ale także program i skrypty. Historie i pytania rozumieją całe osoby, a nie ich ośrodki językowe czy płaty czołowe. Podobnie, jak twierdzi Systems Repy, nie ma znaczenia, że pewna część systemu (tj. człowiek w pokoju chińskim) nie rozumie; to całemu Pokojowi Chińskiemu należy przypisać zrozumienie. Odpowiedzią Searle'a na Odpowiedź Systemu było wyobrażenie sobie, że zapamiętuje program i skrypty oraz wykonuje program, sprawdzając swoją pamięć. Efektem internalizacji programu i skryptów w pamięci jest to, że granica systemu jest teraz taka sama jak granica

ciała Searle'a. Jednak, argumentował Searle, nadal nie zrozumiałby ani słowa po chińsku, mimo że efektem wykonania jego cudownego programu umysłowego byłyby pisane produkty, które rodzimi użytkownicy języka chińskiego uważaliby za nienagannie poprawne. Chińska instrukcja „Krzycz, jeśli chcesz hamburgera" nadal nie dawałaby Searle'owi powodu do krzyku, nawet gdy pisał po chińsku: „Tak, chcę hamburgera". Inną znaczącą odpowiedzią, którą rozważał Searle, jest Odpowiedź Robota. Ta odpowiedź stanowi ważne ustępstwo w stosunku do argumentu Searle'a, a mianowicie, że zwykły komputer nie rozumie, niezależnie od tego, jak dobre są jego werbalne odpowiedzi. Mówi jednak, że gdyby wyjścia komputera zostały wykorzystane do sterowania robotem w taki sposób, aby działania pasowały do słów, cały robot by zrozumiał. W takim przypadku istniałyby analogie percepcji i działania, które łączyłyby słowa z przedmiotami i sytuacjami w świecie. Na przykład komputer, który odpowiedział: „Co powinieneś zrobić, jeśli poczujesz zapach dymu?" z „Opuść pokój" może w ogóle nie zrozumieć. Załóżmy jednak, że robot ma czujnik dymu i załóżmy, że jego zdolność do generowania dobrej odpowiedzi jest połączona z mechanizmami transdukcyjnymi, które również zmuszają go do opuszczenia pomieszczenia po pobudzeniu czujnika dymu. Można by wtedy powiedzieć, że o takim robocie nie tylko dobrze gra, ale też rozumie, co mówi. Wiarygodność tej sugestii staje się tym większa, im szerszy jest zakres nowatorskich reakcji, którym towarzyszą odpowiednie działania, jakie sobie wyobrażamy. Odpowiedzią Searle'a na Odpowiedź Robota było wyobrażenie sobie scenariusza, w którym znajduje się w pokoju chińskim, który znajduje się wewnątrz robota. Postępuje tak jak poprzednio; ale teraz, bez jego wiedzy, wiadomości, które przekazuje z Pokoju Chińskiego, wpływają nie tylko na mowę robota, ale także na jego odpowiednie działania. Searle zauważył, że w tym scenariuszu nadal nie zrozumiałby żadnego ze słów, które przetwarza, i doszedł do wniosku, że robot, którego (nieświadomie) kontroluje, nie ma intencjonalności. Niestety, pomimo poświęcenia części temu, co nazywa „odpowiedzią kombinowaną", Searle tak naprawdę nie omawia wyniku połączenia odpowiedzi systemu i odpowiedzi robota. Kombinacja ta pozwala na zarzut, że granicą właściwą dla osoby rozumiejącej w przypadku robota nie jest człowiek znajdujący się wewnątrz, ale cały robot. Co więcej, internalizujące posunięcie Searle'a w odpowiedzi na Odpowiedź Systemu nie sprawdzi się tutaj; robotyczny odpowiednik internalizacji skryptów i programu wymagałby internalizacji mechanizmów kierujących odpowiednimi działaniami. Na przykład robot sterowany przez program Searle'a może nie tylko dawać werbalne komunikaty „Kiedy baterie się wyczerpują, najlepiej udać się do najbliższej ładowarki", ale także udać się do najbliższej ładowarki, gdy baterie są rozładowane. Brak wyraźnego uwzględnienia przez Searle'a tej możliwości pozostawia jego argumentację otwartą pod zarzutem, że chociaż miał rację, że komputery nie rozumieją, nie wykazał, że odpowiednio zorganizowane roboty nie mogłyby zrozumieć, co mówią. Kwestię, czego potrzeba, aby sztucznie stworzonym słowom nadać autentyczną treść semantyczną, omawiali inni, zwłaszcza Harnad (1990), w ramach wyrażenia „problem uziemienia symboli".

Dzieło Dreyfusa

Hubert Dreyfus (1972/1979) poruszył szereg wpływowych wyzwań stojących przed klasycznym podejściem do sztucznej inteligencji. Możemy zacząć je rozumieć, wracając do naszego przykładu kalkulatorów. Prawdą jest, że takie urządzenia ucieleśniają zasady arytmetyczne; jednak prawdopodobnie nie są (rzeczą) inteligentni. Inteligencji wymagało odkrycie zasad arytmetyki, a kalkulatory w ogóle nie rozpoczęły tego projektu. Można je przekonująco uważać za zwykłe narzędzia, w których przechowujemy pewien produkt naszej inteligencji, i jako takie nie rzucają żadnego światła na to, czego potrzeba, aby faktycznie być inteligentnym. Bardziej złożone urządzenia mogą przechowywać bardziej imponujące produkty naszej inteligencji, ale wniosek, jaki należy wyciągnąć, jest w zasadzie taki sam. Inteligencja wymaga umiejętności ustalenia, jakie zasady należy zastosować – i kiedy można je zastosować – a maszyny, które jedynie przechowują reguły, nie zaczynają tego robić, niezależnie od tego, jak przydatne mogą być jako narzędzia. Wyzwanie to można pogłębić, zauważając,

że inteligencja przejawia się w umiejętności rozpoznawania tego, co jest istotne dla każdego wykonywanego zadania. Ale ogólnie rzecz biorąc, znaczenie zależy od wszystkich cech obecnych (lub nieobecnych!) w sytuacji. Zatem osiągnięcie uznania istotności poprzez zastosowanie reguł do reprezentacji cech wydaje się wymagać wyczerpujących zestawów reguł, które mają zastosowanie do każdej ewentualności. Z jednej strony projekt ten wydaje się mało realny. Z drugiej strony, gdyby dało się to przeprowadzić, można byłoby postawić zarzut, że w powstałym urządzeniu nie została zawarta inteligencja, która w najlepszym przypadku byłaby repozytorium wyników naszej inteligencji. Problem wykorzystania tego, co istotne dla bieżących zadań, nazywany jest czasem „problemem ramowym". Problem ten jest czasami uważany za czysto empiryczny, to znaczy problem, który należy rozwiązać poprzez znalezienie odpowiedniego kompromisu pomiędzy ograniczeniami dotyczącymi zakresu przypadków, do których można zastosować program (co może zmniejszyć rozmiar przestrzeni, którą należy przeszukiwać) odpowiednie dane lub operacje) oraz złożoność programu potrzebnego do odpowiedniego poradzenia sobie z zamierzoną przestrzenią problemową. Wyzwanie filozoficzne jest dodatkiem do trudności empirycznych (które są znaczne); sugeruje, że cała koncepcja podejścia do inteligencji poprzez reprezentacje i reguły nigdy nie zapewni narzędzia wykazującego prawdziwą inteligencję. Zawiera także wyjaśnienie trudności wielokrotnie napotykanych przy próbach „zwiększenia skali" eleganckich rozwiązań stosunkowo prostych problemów, tak aby można je było zastosować do złożonych problemów świata rzeczywistego. Wyjaśnienie jest takie, że wydajność implikuje ograniczenie rozmiaru przestrzeni poszukiwań, ale sukces w zapewnianiu elastyczności wymaga umożliwienia potencjalnego dostępu do dowolnego fragmentu danych. Nie jest oczywiste, w jaki sposób można rozwiązać napięcie między tymi dwoma żądaniami w przypadku problemów występujących w świecie rzeczywistym, w których wiele cech może okazać się istotnych dla skutecznego radzenia sobie. Z pewnych praktycznych powodów problemy te można rozwiązać poprzez postęp sprzętowy w zakresie rozmiaru pamięci i szybkości przetwarzania. Jednakże im bardziej rozwiązania opierają się na tych środkach, tym bardziej prawdopodobne staje się, że musi istnieć alternatywne podejście do inteligencji; bo choć pojemność naszego mózgu jest duża, to prędkość jego przetwarzania na poziomie poszczególnych elementów jest powolna (około 100 kroków przetwarzania na sekundę [Feldman 1985] w porównaniu z milionami w przypadku komputerów). W terminologii wprowadzonej powyżej, im bardziej klasyczne podejścia do sztucznej inteligencji opierają swoje ulepszenia wydajności na postępie w oprogramowaniu ardware, tym bardziej wygląda na to, że co najwyżej dadzą zwykłą sztuczną inteligencję. Ta refleksja sugeruje, że dążenie do sztucznej inteligencji premium może być przydatne nawet w zastosowaniach praktycznych, i skłoniło wielu członków społeczności AI do poszukiwania podejść bardziej inspirowanych biologią.

Koneksjonizm

Urządzenia koneksjonistyczne (równoległe procesory rozproszone, sztuczne sieci neuronowe) zazwyczaj składają się z jednostek, których moc wyjściowa przekazywana do innych jednostek jest funkcją sumy ważonych danych wejściowych otrzymywanych od innych jednostek. Ważony wkład do jednostki B to wynik jednostki A pomnożony przez wagę połączenia (dodatniego lub ujemnego) pomiędzy jednostką A i jednostką B. Sieci jednostek charakteryzują się częściowo wzorcami połączeń. Zbadano wiele takich wzorców, w tym sieci ściśle ze sprzężeniem zwrotnym i sieci z połączeniami bocznymi („feedsideways") i rekurencyjnymi (ze sprzężeniem zwrotnym). Urządzenia koneksjonistyczne charakteryzują się ponadto regułą znajdowania zestawu wag połączeń, które dadzą wzorce na ich jednostkach wyjściowych, które są odpowiednie dla każdego wzorca na ich jednostkach wejściowych. W wielu przypadkach różnice między rzeczywistymi wynikami a prawidłowymi wynikami wzorców w zestawie treningowym są wykorzystywane do generowania sygnału błędu, który z kolei służy do przyrostowego dostosowywania wag. Po wielu takich cyklach regulacji osiągane są prawidłowe pary wejść i wyjść. Urządzenia koneksjonistyczne mają kilka właściwości, które wzbudziły

zainteresowanie. Na przykład generalizują w tym sensie, że po wyszkoleniu nowy wzorzec wejściowy, który jest podobny do wzorca szkoleniowego P1, wygeneruje wynik podobny do wyniku P1. Ich „pamięć" jest „adresowalna treściowo" – to znaczy wzorzec wejściowy wraz z istniejącym zestawem wag bezpośrednio generuje wynik i nie ma procesu „wyszukiwania" informacji istotnych dla relacji wejście–wyjście. Ulegają degradacji z gracją – to znaczy uszkodzenie niektórych jednostek sieci nie niszczy natychmiast relacji wejście-wyjście. Zamiast tego istnieje zakres, w którym uszkodzenie powoduje, że uogólnienie stopniowo się pogarsza, ale nie jest całkowicie bezużyteczne.

Wyniki przypominające reguły bez reguł?

Wymienione właśnie właściwości wzbudziły zainteresowanie po części dlatego, że przypominają nasze własne właściwości psychologiczne, co sugeruje, że badania koneksjonistyczne mogą zapewnić pewien wgląd w sztuczną inteligencję premium. Sugestia ta występuje w przypadku kolejnej i nieco kontrowersyjnej właściwości niektórych urządzeń koneksjonistycznych; mianowicie wydają się być w stanie zapewnić przypominające reguły relacje między wzorcami wejściowymi i wyjściowymi bez posiadania odpowiadających im wewnętrznych reprezentacji reguł. Na przykład wydaje się, że kiedy dzieci uczą się, jak tworzyć czas przeszły w języku angielskim, uczą się zasady, która ma zastosowanie do większości czasowników: „dodaj „-ed". Jednakże w szeroko omawianym eksperymencie Rumelhart i McClelland (1986) wytrenowali koneksjonistyczne narzędzie umożliwiające powiązanie fonetycznych reprezentacji czasowników z fonetycznymi reprezentacjami ich form w czasie przeszłym. Uwzględniono zarówno czasowniki regularne, jak i nieregularne, a wyniki treningu były dość dobrze uogólnione w przypadku obu. Odtworzono nawet efekt interferencji na czasy przeszłe w przypadku czasowników nieregularnych, który obserwuje się u dzieci i który zwykle przypisuje się nadmiernemu stosowaniu przez nie reguły dla czasów przeszłych czasowników regularnych. Ale trening polegał po prostu na skojarzeniu reprezentacji fonetycznych czasu teraźniejszego i przeszłego. Wyniki można opisać w kategoriach reguł, ale w systemie, który wygenerował wyniki, nie było reprezentacji reguł. Jednakże eksperyment ten był krytykowany z kilku powodów. Najbardziej uogólniona z tych krytyk dotyczy zakresu, w jakim oczywiste Powodzenie w dostarczaniu wyników przypominających reguły bez wyraźnych reguł zależy od konkretnych cech przebiegu szkolenia. Jeśli okaże się, że wyniki przypominające reguły zależą od cech zbioru uczącego, które nie są powszechne w ludzkim doświadczeniu, wówczas ich znaczenie dla możliwości sztucznej inteligencji premium zostanie znacznie osłabione.

Systematyczność

Fodor i Pylyshyn wysunęli linię krytyki koneksjonizmu opartą na ich koncepcji systematyczności. Języków naturalnych uczy się w systemach zdań. Na przykład, jeśli ktoś potrafi zrozumieć i zastosować słowa „Jan kocha Marię", można także zrozumieć i zastosować słowa „Mary kocha Jana", „Tom kocha Jane" i tak dalej. Systematyczność można łatwo wytłumaczyć w kategoriach reguł mających zastosowanie do reprezentacji strukturalnych – na przykład znaczenie „X kocha Y" jest takie, że pierwszy wymieniony element kocha drugi wymieniony element. Jeśli jednak zdania i znaczenia są jedynie powiązane w wyszkoloną sieć, nie ma powodu oczekiwać, że uda się znaleźć systematyczność. Koneksjoniści mogą odpowiedzieć, że istnieją pewne rodzaje sieci, które zapewnią systematyczność. Tego rodzaju odpowiedź rodzi jednak pytanie, czy atrakcyjne właściwości sieci koneksjonistycznych rzeczywiście przyczyniają się do naszego zrozumienia, w jaki sposób wytwarzana jest inteligencja, czy też, jak sugeruje Fodor, sieci koneksjonistyczne są jedynie w stanie wdrożyć klasyczną architekturę – w takim przypadku jest to właściwie to drugie wyjaśnia, w jaki sposób udaje nam się być inteligentnymi. Jeśli to prawda, inteligencję premium można w zasadzie jednak wdrożyć w urządzeniu niekoneksjonistycznym.

Realizm psychologiczny

Można oczywiście rozróżnić pytania: „Czy urządzenia koneksjonistyczne są etapami na drodze prowadzącej do sztucznej inteligencji premium?" oraz „Czy urządzenia koneksjonistyczne mogą zostać wykorzystane do zapewnienia regularnej sztucznej inteligencji?" Negatywna odpowiedź na pierwsze pytanie byłaby zgodna z możliwością powstania robota napędzanego mózgiem koneksjonistycznym, elastycznie reagującego na przeszkody i wykazującego stałą tendencję do osiągania stanów docelowych. Jednak część inspiracji dla koneksjonizmu wynika z pomysłu, że skoro nasze mózgi prawdopodobnie czynią nas inteligentnymi, urządzenie inspirowane mózgiem powinno również być w stanie zapewnić najwyższej jakości sztuczną inteligencję. Dopóki istnieją wątpliwości, czy koneksjonistyczne urządzenia zdołają uchwycić sposób, w jaki pracujemy, inspiracja ta słabnie i dość pożądane jest alternatywne wsparcie dla prawdopodobieństwa zapewnienia przez nie nawet zwykłej sztucznej inteligencji. Wątpliwości tego rodzaju dotyczą wykorzystania sygnałów błędu (różnic między poprawnym wyjściem a rzeczywistym wyjściem jednostki lub zestawu jednostek) w celu znalezienia zestawu wag, który będzie przydatny w danym celu. Istnieje algorytm (powszechnie znany jako propagacja wsteczna), którego skuteczność można wykazać i który jest często stosowany w badaniach koneksjonistycznych. Niestety nie jest jasne, w jaki sposób mózg mógłby zastosować ten algorytm do regulacji połączeń synaptycznych między neuronami. Alternatywnym podejściem do dostosowywania ciężaru jest uczenie się przez wzmacnianie. Podstawową ideą uczenia się przez wzmacnianie jest to, że nikt nie musi znać prawidłowych rozwiązań problemu, jakie działania podjąć; wybory dotyczące działań mogą być kształtowane przez (dodatnią lub negatywną) wartość konsekwencji ostatnio podjętych działań dla podmiotu. Wydaje się to oczywiste ,że możemy się w ten sposób uczyć, a bardzo obiecującą sugestią jest zaprojektowanie urządzeń, które mogą uczyć się również poprzez wzmacnianie. Jednakże ci, którzy wzięli sobie do serca krytykę GOFAI dokonaną przez Dreyfusa, mogą mieć analogiczne wątpliwości dotyczące uczenia się przez wzmacnianie. Prace w tej dziedzinie skupiają się obecnie na małych problemach i nie jest jasne, czy wyniki można skalować, aby sprostać złożoności świata rzeczywistego. Zachowania można dobrze zdefiniować i łatwo rozróżnić. Inteligencja natomiast (czy to zwykła, czy premium) prawdopodobnie będzie miała do czynienia z działaniami, których właściwa klasyfikacja zależy od okoliczności – na przykład tym samym zachowaniem podczas biegania może być ucieczka lub pościg, w zależności od tego, czy drapieżnik znajduje się za biegaczem, czy ofiara jest przed nim. Może się zdarzyć, że praca nad problemami, które pozwalają na łatwą klasyfikację działań, omija istotny element inteligencji; mianowicie uczenie się, jak okoliczności wpływają na znaczenie zachowania. Kolejnym problemem w ocenie bieżących prac nad uczeniem się przez wzmacnianie jest to, że wiele schematów wymaga aktualizacji wartości nie tylko działań, ale stanów będących częściami sekwencji stanów prowadzących do działań. Istnieją algorytmy, które to umożliwiają, ale wymagają one znacznych obliczeń i nie jest oczywiste, w jaki sposób one lub ich odpowiedniki mogłyby zostać wykonane w mózgu. W przypadku zwykłej sztucznej inteligencji nie stanowi to problemu. Jednakże w zakresie, w jakim badania nad uczeniem się przez wzmacnianie wyrzekną się wiarygodnego powiązania z sztuczną inteligencją premium, tracą poparcie dla argumentu, że skoro my, ludzie, wykazujemy umiejętność uczenia się przez wzmacnianie, (regularną) sztuczną inteligencję będzie można osiągnąć dzięki (obecnym) podejściu do uczenia się przez wzmacnianie.

Reprezentacja

Atrakcyjne właściwości koneksjonizmu zależą od zastosowania rozproszonych reprezentacji; to znaczy reprezentacje zależne od wzorca aktywacji w więcej niż jednej jednostce. „Wiedza" urządzenia koneksjonistycznego jest również rozłożona na ciężary jego połączeń. Wyniki zależą od połączonych efektów wielu czynników wagi, a każda waga ma wpływ na wiele wyników. Rozproszony charakter reprezentacji i „wiedzy" w urządzeniach koneksjonistycznych prowadzi do dwóch rodzajów pytań.

Ramsey, Stich i Garon (1990) zauważyli, że rozproszony charakter „wiedzy" sieci utrudnia zrozumienie, w jaki sposób sieć koneksjonistyczna mogłaby kiedykolwiek modelować pozornie oczywistą prawdę, że ludzie mogą działać z tego, a nie innego powodu, nawet jeśli mają przekonania i pragnienia odpowiadające dwóm lub większej liczbie powodów tego działania. Jeśli bowiem odpowiednie przekonania są przechowywane w jednej sieci, połączenia uziemiające oba miałyby wpływ na dowolny wynik. Można oczywiście utrzymywać, że różne przekonania są przechowywane w różnych sieciach; ale ogólnie zastosowanie tej strategii oznaczałoby rezygnację z domniemanych zalet sieci koneksjonistycznych. Koneksjoniści mogą jednak odpowiedzieć, że rozróżnienia między przyczynami wyniku można odpowiednio dokonać, zwracając uwagę na wszelkie różnice, które dają nam podstawę do przekonania, że działanie zostało podjęte z jednego, a nie drugiego, z dwóch dobrych powodów (Robinson 1995). Może się na przykład zdarzyć, że w danej sytuacji obecny był (lub był istotny) tylko jeden z dwóch możliwych danych wejściowych. Jeśli to I1 powoduje O, przyczyna związana z I1 może być przyczyną wyjścia, nawet jeśli sieć dałaby O, gdyby wejście I2 było obecne (lub bardziej znaczące). Kontrowersyjne jest jednak to, czy można zapewnić analogi tej strategii dla wszystkich istotnych przypadków. Drugie, szeroko dyskutowane zagadnienie dotyczy przetwarzania reprezentacji w systemach koneksjonistycznych. Istnieje kilka pomysłowych schematów tworzenia i przechowywania informacji w urządzeniach koneksyjnych. (Patrz np. Pollack 1988, 1990; Smolensky i Legendre 2006). Jeśli jednak przechowywane reprezentacje muszą zostać odkodowane (tj. odzyskiwane i reprezentowane osobno), aby można je było wykorzystać w poznawczo użytecznym przetwarzaniu, wówczas koncepcja, że koneksjonizm tworzy charakterystyczny wkład w naszym rozumieniu poznania budzi pewne wątpliwości. W genialnym eksperymencie Chalmers (1993) wykazał, że reprezentacje koneksjonistyczne można przetwarzać bez dekodowania. W eksperymencie tym uczestniczyły dwie sieci. Pierwszą sieć wyszkolono w zakresie tworzenia skompresowanej reprezentacji zdań w głosie czynnym i zdań w stronie biernej oraz dekodowania tych reprezentacji w ich oryginalne pełne zdania. Drugą sieć przeszkolono w zakresie przekształcania skompresowanych reprezentacji zdań w głosie czynnym w skompresowane reprezentacje ich odpowiedników w głosie biernym. Kiedy te ostatnie skompresowane reprezentacje zostały zdekodowane przez pierwszą sieć, wynikiem był poprawny odpowiednik strony biernej odpowiedniego zdania w głosie czynnym. Co najciekawsze, ten sam wynik uzyskano dla skompresowanych reprezentacji, które nigdy nie zostały zaprezentowane drugiej sieci podczas jej fazy uczenia. W rezultacie druga sieć była w stanie zastosować to, czego „nauczyła się" na temat relacji głosu aktywnego/biernego do nowych (tj. niewidzianych) przypadków; i był w stanie to zrobić bez uprzedniego dekodowania skompresowanych reprezentacji głosu aktywnego. Jednakże, jak wyraźnie zauważył Chalmers, wynik ten nie był realistyczny z psychologicznego punktu widzenia. Zbiór uczący stanowił ponad połowę całego zbioru przypadków, a wszystkie zdania miały tę samą prostą formę (podmiot, czasownik, dopełnienie). Tak więc, choć eksperyment Chalmersa dostarcza dowodu na możliwość użytecznego przetwarzania bez dekodowania, otwartym pytaniem pozostaje, czy można znaleźć psychologicznie realistyczne przykłady. Z szerokiej perspektywy GOFAI i koneksjonizm zgadzają się, że inteligencja obejmuje przetwarzanie reprezentacji. Ich różnica zdań dotyczy charakteru tego przetwarzania i w świetle trudności związanych z obydwoma podejściami być może nie jest zaskakujące, że podejrzenia padły na samą ideę reprezentacji.

Teoria układów dynamicznych

Teoria systemów dynamicznych (DST) postrzega poznanie jako zależne od ciągłej interakcji czynnika poznawczego z otoczeniem. Inteligentne działanie nie powstaje w wyniku najpierw reprezentowania środowiska, a następnie wykonywania procesów na podstawie tej reprezentacji. Zamiast tego nakłady środowiskowe bezpośrednio napędzają działania, które oddziałują z powrotem na środowisko, powodując nowe wejścia i nowe reakcje. Inteligencja wynika z projektu, który może wykorzystywać informacje obecne w środowisku bez uprzedniego przekształcania tych informacji w wewnętrzną

reprezentację. Kluczową cechą układów dynamicznych jest ich związek z czasem. Jeśli urządzenie działa w oparciu o reprezentacje, jego operacje mogą zmieniać się w czasie, pod warunkiem jedynie, że reprezentacja działania musi zostać dostarczona na czas, aby działanie było przydatne. Jeśli przetwarzanie urządzenia jest ściśle powiązane z zewnętrznymi danymi wejściowymi, a nie z ich reprezentacjami, wówczas czas procesu zależy od czasu nadejścia tych danych wejściowych i nie ma arbitralności w przebiegu czasowym przetwarzania. Idee te zilustrowano na przykładzie gubernatora Watta. Urządzenie to łączy wrzeciono z wałem napędowym maszyny parowej. Wrzeciono podtrzymuje parę obciążników, które obracają się wraz z wrzecionem i które są zamontowane w taki sposób, że zmieniają swoją wysokość w miarę zmiany prędkości obrotowej silnika w zależności od odległości od osi wrzeciona. Ramiona podtrzymujące obciążniki są mechanicznie połączone z zaworem, który zmienia ciśnienie pary odwrotnie proporcjonalnie do wysokości obciążników. Efektem takiego rozwiązania jest utrzymanie prędkości obrotowej silnika w wąskim zakresie, nawet przy zmiennym obciążeniu wału napędowego. Van Gelder zasugerował, że jeśli próbujemy zrozumieć poznanie, gubernator Watta jest lepszą inspiracją niż komputery. Kluczowym punktem kontrastu jest to, że system powiązań gubernatora spełnia swoją użyteczną funkcję, nie posiadając żadnej części, która stosuje reguły do reprezentacji. Prędkość wału napędowego jest mechanicznie połączona z regulatorem, który z kolei jest mechanicznie sprzężony z zaworem. Rezultatem jest układ dynamiczny, który nie wymaga obliczania odpowiedniego stanu zaworu na podstawie zapisu prędkości wału napędowego. Przykłady robotów Brooksa (1991) są bardziej złożone niż przykłady gubernatora Watta, ale inspiracja – widoczna w jego tytule „Inteligencja bez reprezentacji” – jest taka sama. Brooks uważa nawigację za istotny problem, który został rozwiązany w czasie ewolucyjnym, zanim wykształciła się umiejętność wykonywania jawnych obliczeń. Jego praca sugeruje, że najlepiej zrozumiemy inteligencję, konstruując urządzenia, które dynamicznie łączą dane wejściowe z działaniami, to znaczy urządzenia, które zawsze podlegają ograniczeniom podobnym do tych, które występowały na przestrzeni ewolucji. Brooks zdecydowanie odrzuca jakąkolwiek koncepcję, w której istnieje podział pracy pomiędzy urządzeniami ucieleśniającymi inteligencję i urządzeniami zamieniającymi wynik w działania. Problemy z przekształcaniem bodźców zmysłowych na reprezentacje komputerowe i reprezentacje komputerowe na dane wyjściowe silnika są w najlepszym wypadku ogromne (a w najgorszym – nierozwiązywalne). Należy zatem porzucić takie podejście na rzecz urządzeń, w których na każdym etapie wbudowane jest połączenie wejścia sensorycznego z wyjściem silnika.

Pytania dotyczące czasu letniego

Bagatelizowanie reprezentacji w naturalny sposób skupia uwagę na tym, co należy uznać za „reprezentację”. Najsłabsza interpretacja tego terminu utożsamiałaby go ze „śledzeniem” – czyli wysoką korelacją między występowaniem stanów układu uznawanych za reprezentacje, a obecnością obiektów lub właściwości, które uważa się za reprezentowane. Znacznie bogatsza koncepcja „reprezentacji” wymaga modelu wewnętrznego, który można wykorzystać do przewidywania przyszłych stanów na podstawie obecnych danych wejściowych. Wydaje się jasne, że systemy dynamiczne mogą działać bez reprezentacji w tym silniejszym znaczeniu. Nie jest jednak tak jasne, czy urządzenia tego rodzaju stosowane w teorii układów dynamicznych rezygnują z „reprezentacji” w sensie zwykłego śledzenia. Na przykład wysokość ciężarków w regulatorze Watta można uznać za reprezentującą prędkość wału napędowego. Kolejne pytanie (również postawione przez Bechtela 1998) dotyczy zakresu, w jakim teoria układów dynamicznych oferuje charakterystyczne wyjaśnienie poznania, jak w przeciwieństwie do odrębnego rodzaju opisu tego, co robią systemy poznawcze. Aby to zilustrować: Zależność pomiędzy wysokością ciężarków w regulatorze Watta a prędkością silnika można elegancko opisać za pomocą równań różniczkowych. Możemy jednak zadać sobie pytanie, jak działa urządzenie i udzielić odpowiedzi, wskazując na wrzeciono, zawór, wał napędowy i mechaniczne połączenia między nimi. Jeśli to rozróżnienie między opisem a mechanizmem wyjaśniającym można

zastosować do systemów poznawczych, wówczas wkład teorii systemów dynamicznych, choć cenny na poziomie opisowym, może nie zapewniać radykalnego odejścia na poziomie wyjaśniania poznania. Wreszcie problem teorii systemów dynamicznych wynika z obserwacji, że wiele przypadków inteligentnego działania w dużym stopniu opiera się na pamięci. Wydaje się, że użycie pamięci wymaga reprezentacji, a te reprezentacje muszą mieć wpływ na zachowanie niezależnie od czasu, w którym reprezentacja pamięci została utworzona. Badania nad sterownikami robotów działającymi bez pamięci wykazały zaskakujące możliwości rozwiązywania problemów; niemniej jednak nie jest prawdopodobne, aby istniały urządzenia, które będą powszechnie akceptowane jako wykazujące inteligencję (rzeczy), ale nie opierające się na pamięci. Oczywiście, reprezentacje pamięci można włączyć do DST. Nie jest jednak jasne, jak można to zrobić, nie cofając nas do wcześniej omówionych pytań dotyczących tego, w jaki sposób reprezentacje mogą być przetwarzane w celu uzyskania inteligentnych wyników.