Przegląd Filozoficzny – Nowa Seria
R. 22: 2013, Nr 2 (86), ISSN 1230–1493
DOI: 10.2478/pfns-2013-0032

Jarek Gryz

Gdzie jesteś, HAL?

Słowa kluczowe: *filozofia umysłu, sztuczna inteligencja, kognitywistyka, maszyna myśląca, reprezentacja wiedzy, robotyka*

1. Początki

Umysł to komputer. Teza ta, która nawet w tak niedoprecyzowanej formie jest dla wielu z nas nie do przyjęcia, legła u podstaw *sztucznej inteligencji*[1], jednej z najbardziej fascynujących i kontrowersyjnych dziedzin nauki zapoczątkowanych w ubiegłym stuleciu. Była to dziedzina, która bodaj jako pierwsza wyodrębniła się z informatyki jako osobna poddziedzina, ale miała ambicje daleko poza informatykę wykraczające. Przyciągnęła ona najtęższe umysły matematyki, ekonomii, psychologii i filozofii, ale skonsumowała też ogromne (prawdopodobnie największe w informatyce) fundusze badawcze. Jej zadufanie i bombastyczne obietnice z pierwszych lat istnienia z czasem ustąpiły przesadnej wręcz skromności i chęci zdegradowania dziedziny przez nazywanie jej *racjonalnością obliczeniową* zamiast sztuczną inteligencją (Russell i Norvig 2010).

W artykule tym postaram się prześledzić dyskusję metodologiczną towarzyszącą sztucznej inteligencji od początku jej istnienia aż po dziś. Z konieczności będzie to opis bardzo skrótowy, bo w samej tylko filozofii wyprodukowano dotąd blisko 3000 publikacji na temat sztucznej inteligencji (Dietrich 2012). W istocie najbardziej interesować nas tu będzie właśnie filozoficzna dyskusja nad podstawami sztucznej inteligencji, albowiem kluczowe w tej dziedzinie

[1] Termin *sztuczna inteligencja* jest dwuznaczny, bo używa się go zarówno do określania dziedziny badawczej, jak i jej potencjalnego produktu. Aby tej dwuznaczności uniknąć, będziemy używać tego terminu tylko w pierwszym znaczeniu, rezerwując dla drugiego pojęcia określenie „maszyna myśląca".

kwestie, takie jak relacja mózgu i umysłu czy racjonalność ludzkiego działania i myślenia, dyskutowane były przez filozofów na długo zanim zbudowano pierwszy komputer. Chcielibyśmy również spróbować określić różnicę czy też relację między sztuczną inteligencją a kognitywistyką (będzie to o tyle trudne, że jak się zdaje, sztuczna inteligencja *już* nie wie, a kognitywistyka *jeszcze* nie wie, czym jest).

Pierwszy amerykański komputer, ENIAC, został uruchomiony w 1945 roku. Jego zastosowania były czysto wojskowe i dotyczyły symulacji wybuchu i potencjalnych zniszczeń spowodowanych przez projektowaną wówczas bombę wodorową. W powszechnym mniemaniu komputer to było po prostu szybkie liczydło (nawet nie kalkulator, bo tych oczywiście wtedy nie używano), zdolne wyłącznie do manipulowania liczbami. Warto o tym pamiętać, bo trzeba było nie lada geniuszu, by wyobrazić sobie inne zastosowania dla owego li tylko „liczydła". Geniuszem tym wykazał się Herbert Simon, który pracował wówczas nad komputerową symulacją obrony powietrznej w RAND Corporation:

> Miałem już wówczas styczność z komputerem (...), ale laboratorium obrony powietrznej to było olśnienie: mieli to wspaniałe urządzenie do symulacji map na maszynach liczących. (...) Nagle stało się dla mnie jasne, że nie musimy ograniczać się do manipulowania liczbami – można wyliczyć pozycję, którą chciałeś, i wydrukować to potem na papierze (McCorduck 2004: 148).

Tak oto narodził się paradygmat komputera jako maszyny do przetwarzania informacji: komputer operuje na symbolach odnoszących się do obiektów istniejących realnie. Stąd był już tylko krok do funkcjonalnego zidentyfikowania komputera i ludzkiego umysłu. Simon wspomina:

> Kiedy po raz pierwszy zauważyłem, że można postrzegać komputer jako urządzenie do przetwarzania informacji, a nie tylko liczb, wtedy metafora umyłu jako czegoś, co zbiera przesłanki, przetwarza je i generuje konkluzje, zaczęła przeistaczać się w przekonanie, że umysł to program, który dostaje jakieś dane na wejściu, przetwarza je i generuje dane na wyjściu. Jest więc bezpośredni związek, w pewnym sensie bardzo prosty, między tym wcześniejszym poglądem, że umysł to maszyna logiczna, a poglądem późniejszym, że umysł to komputer (McCorduck 2004: 151).

Zanim spróbujemy przeanalizować argumenty logiczne, które doprowadziły do wniosku, że umysł to komputer, zwróćmy uwagę na pewne istotne argumenty *psychologiczne*. Dotyczą one kilku niezwykle – jak na tamte czasy – spektakularnych osiągnięć inżynierskich dokonanych w sztucznej inteligencji w latach 50.

Niemal od pierwszej chwili, kiedy odkryto, że komputer może przetwarzać dowolne symbole (a nie tylko liczby), usiłowano stworzyć program, który pokazałby możliwości maszyny w tej dziedzinie. Oczywistym zastosowaniem były szachy: gra, której reguły łatwo jest opisać w języku symbolicznym, a w której – jak się początkowo wydawało – szybkość przetwarzania symboli (pozycji na szachownicy i ich „wartości") miała kluczowe znaczenie dla pokazania przewagi maszyny nad człowiekiem. Wnet jednak okazało się, że „bezmyślne" przeszukiwanie wszystkich sekwencji ruchów na szachownicy dla znalezienia takiej, która prowadziłaby nieodzownie do wygranej, jest nierealne. Claude Shannon oszacował liczbę możliwych posunięć na 10^{120}, co znaczy, że przy weryfikacji jednej sekwencji w ciągu jednej milionowej sekundy pierwszy ruch na szachownicy wykonany byłby po 10^{95} latach[2]. Herbert Simon i współpracujący z nim wówczas Allen Newell nie zamierzali oczywiście tak długo czekać. Zmienili więc dziedzinę i postanowili napisać program, który dowodziłby twierdzeń logiki. Choć złożoność obliczeniowa takiego programu jest dużo mniejsza niż gry w szachy, to i w tym przypadku konieczne było sformułowanie reguł i metod efektywnego poszukiwania dowodu, a nie proste weryfikowanie, czy dowolnie wygenerowany ciąg znaków spełnia wymogi dowodu w sensie logicznym. Program, który w ten sposób powstał, Logic Theorist (Newell and Simon 1956), był więc w pewnym sensie „kreatywny", bo generował dowody, których jego autorzy się nie spodziewali. Innymi słowy, choć program spełniał wymagania postawione przez jego autorów, jego zachowania nie można było (przynajmniej w łatwy sposób) przewidzieć. Sukces programu był spektakularny: dowiódł on 38 z 52 twierdzeń drugiego rozdziału *Principia Mathematica* Russella i Whiteheada. Co więcej, dowód twierdzenia 2.85 sformułowany przez Logic Theorist okazał się bardziej elegancki niż ten z *Principia.* Russell był pod wielkim wrażeniem owego sukcesu, ale już „The Journal of Symbolic Logic" odmówił publikacji dowodu, którego autorem była maszyna.

Tak oto Newell i Simon pokazali, że maszyna zdolna jest do wykonywania zadań, które wymagały wyobraźni, inteligencji i kreatywności, a więc cech uważanych powszechnie za zastrzeżone dla człowieka. Nic zatem dziwnego, że w styczniu 1956 roku Simon ogłosił, iż stworzona została pierwsza w historii maszyna myśląca. Nie był w tym przekonaniu odosobniony. Pod koniec lat 50. stworzono jeszcze dwa programy komputerowe, które przewyższały możliwości przeciętnego człowieka w rozumowaniu abstrakcyjnym. Arthur Samuel pracował już od końca lat 40. nad programem do gry w warcaby (Samuel 1959).

[2] Zwycięstwo *Deep Blue* nad szachistą Kasparowem w 1997 roku było możliwe między innymi dlatego, że uniknięto tam prostego przeszukiwania przestrzeni możliwych ruchów (Campbell, Hoane, Hsu 2002).

Choć warcaby są grą prostszą niż szachy, a przestrzeń możliwych do wykonania sekwencji ruchów jest dużo mniejsza, nie jest to jednak gra trywialna. Co więcej, program napisany przez Samuela potrafił się uczyć, a więc zawierał kolejny element myślenia, zastrzeżony – jak się dotąd wydawało – wyłącznie dla ludzi. Kiedy trafiał na pozycję, którą napotkał już w poprzednich grach, oceniał jej wartość na podstawie rezultatów osiągniętych poprzednio, zamiast szacować tę wartość „w ciemno". W 1961 roku program Samuela zwyciężał już rutynowo z mistrzami warcabów. Sukcesem mógł się również pochwalić Herbert Gelernter, którego program (Gelernter 1959) potrafił skonstruować nowe, nietypowe dowody twierdzeń geometrii dwuwymiarowej.

Te wczesne, niewątpliwe sukcesy sztucznej inteligencji skłoniły wielu ówczesnych naukowców zarówno do przeceniania własnych osiągnięć, jak i do stawiania hurra-optymistycznych prognoz na przyszłość.

> Powiem wprost, choć będzie to może dla niektórych szokiem, że są już na świecie maszyny, które myślą, uczą się i tworzą. Co więcej, ich umiejętności rozwijają się na tyle szybko, że w niedalekiej przyszłości będą rozwiązywać problemy, do których dotąd potrzebny był ludzki umysł. (...) Zaprojektowaliśmy program komputerowy zdolny do myślenia nienumerycznego i tym samym rozwiązaliśmy odwieczny problem relacji ciała i umysłu (Simon and Newell 1958: 8).

> W ciągu dwudziestu lat maszyny będą w stanie wykonać każdą czynność, którą dziś wykonuje człowiek (Simon 1965: 96).

Oczywiście badacze sztucznej inteligencji zdawali sobie sprawę, że z faktu, iż *pewne* elementy ludzkiego myślenia mogą być imitowane czy realizowane przez komputer, nie wynika, że dotyczy to elementów *wszystkich*. Potrzebny był więc argument pokazujący, co dla myślenia jest istotne, a co jest tylko przypadkowe. Przypadkowe byłoby więc na przykład fizyczne umiejscowienie procesów myślowych w mózgu; zastąpienie jednego czy wręcz wszystkich neuronów przez elementy mechaniczne (takie jak obwody elektryczne) o tej samej funkcji nie powinno mieć wpływu na procesy myślowe. W istocie teza taka stawiana była niemal natychmiast po pojawieniu się pierwszych komputerów. W artykule napisanym w 1947 roku, a opublikowanym dopiero po 22 latach, Alan Turing pisał: „Obwody elektryczne używane w komputerach wydają się mieć własności istotne dla neuronów. Zdolne są one do przesyłania i przechowywania informacji" (Turing 1969). W podobnym duchu wypowiadał się John von Neumann, który definiując obowiązującą do dziś architekturę systemu komputerowego, używał terminów takich jak „pamięć" czy organy „zarządzania". Omawiając części komputera, pisał również: „Trzy odrębne części (...) odpowiadają neuronom asocjacyjnym w ludzkim systemie ner-

wowym. Pozostaje nam omówić odpowiedniki neuronów sensorycznych, czyli dośrodkowych, i motorycznych, czyli odśrodkowych" (Goldstine 1972).

Co zatem stanowi o istocie myślenia i co pozwala nam ignorować konkretną realizację procesów myślowych? Newell i Simon sformułowali to jako słynną hipotezę systemu symboli: „Fizyczny system symboli to warunek konieczny i wystarczający dla inteligentnego działania" (Newell i Simon 1976: 116). Manipulacja czy też obliczanie (ang. *computation*) przy użyciu tych symboli to właśnie myślenie. Zenon Pylyshyn, jeden z obrońców tej hipotezy, pisze:

> Idea procesu myślowego jako obliczania to ważka hipoteza empiryczna. (...) Umysł jest (...) zajęty szybkim, w dużej mierze nieświadomym przeszukiwaniem, zapamiętywaniem i rozumowaniem, czyli ogólnie mówiąc manipulowaniem wiedzą. (...) [Ta wiedza] zakodowana jest jako własności mózgu w ten sam sposób, jak zakodowana jest semantyczna zawartość reprezentacji w komputerze – przez fizycznie urzeczywistnione struktury symboli (Pylyshyn 1984: 55, 193, 258).

Warto zauważyć, że w powyższej wersji hipotezy (tzw. wersji silnej) ludzki umysł *musi* być takim systemem symboli (w przeciwnym razie należałoby mu bowiem odmówić możliwości myślenia). W takiej wersji zbudowanie maszyny myślącej przy użyciu komputera jest oczywiście teoretycznie wykonalne, z tej prostej przyczyny, że ludzki umysł to po prostu komputer[3]. Wersja silna hipotezy ma jeszcze jedną istotną konsekwencję, a mianowicie, że badając działania komputera (czy ściśle rzecz biorąc zainstalowanego w nim programu), możemy dowiedzieć się czegoś nowego na temat działania mózgu. Na założeniu tym ufundowana została kognitywistyka (Thagard 2008). Zauważmy na koniec, że dla uzasadnienia celów sztucznej inteligencji wystarczy słaba wersja hipotezy, mianowicie, że manipulacja symboli jest wystarczająca, ale niekonieczna dla myślenia.

2. Rozwój

Lata 60. i 70. to złoty wiek sztucznej inteligencji. Gwałtownie wzrosła ilość badaczy zajmujących się tą dziedziną, a także funduszy przeznaczonych na badania. Większość rozwiązań algorytmicznych i systemowych, które weszły do kanonu dziedziny, pochodzi właśnie z tamtych czasów. Co więcej, sukcesy sztucznej inteligencji przestały już być tylko wewnętrzną sprawą naukowców. Media zapowiadały cywilizacyjną rewolucję, która lada dzień dokonać się miała

[3] W sztucznej inteligencji przyjęło się używać terminu *komputer* czy *maszyna* tam, gdzie naprawdę chodzi o *program* (podobne rozróżnienie czyni się między „mózgiem" a „umysłem"). Oczywiście trudno byłoby zmienić w tej chwili uzus językowy i mówić o *programie myślącym*, ale warto o tym niełądzie pamiętać.

za sprawą budowanych właśnie maszyn myślących (Darrach 1970). Ale za tą efektowną fasadą pojawiły się też pierwsze pęknięcia. Zdawano sobie już wcześniej sprawę z ułomności i ograniczeń budowanych systemów, ale dopiero teraz zaczęto wyraźnie rozpoznawać problemy. Jak się później okazało, wielu z nich nie dawało się rozwiązać za pomocą usprawnień technologicznych, bo wynikały one z dokonanych wcześniej założeń metodologicznych czy też filozoficznych, na których opierała się sztuczna inteligencja. Zanim przejdziemy do analizy tych problemów, przedstawmy dwa projekty z zakresu robotyki z tamtego okresu: SHRDLU i Shakey. Wówczas oba te projekty wydawały się sukcesami, a dziś służą raczej do ilustracji problemów, jakie stają przed sztuczną inteligencją.

Celem SHRDLU było zbudowanie robota, który potrafiłby porozumiewać się z człowiekiem, rozumować i przeprowadzać proste działania. I tak SHRDLU dysponował chwytakiem, przy pomocy którego mógł on manipulować rozrzuconymi na stole klockami o różnych kolorach i kształtach. W istocie zarówno środowisko, jak i sam robot były tylko symulacjami: badaczy nie interesowało w tym przypadku rozwiązywanie technologicznego problemu konstrukcji mechanicznego robota, a tylko jego możliwości rozumowania. Komunikacja między takim „robotem" a człowiekiem odbywała się przy pomocy klawiatury do zadawania pytań i wydawania poleceń, oraz monitora, na którym pokazany był stół z klockami i odpowiedzi robota. Kiedy wydawano robotowi polecenie ułożenia klocków w jakiejś konfiguracji, SHRDLU wymyślał plan działania i wirtualnie go wykonywał. SHRDLU potrafił zrozumieć język potoczny i do pewnego stopnia rozumiał swoje motywy. Oto przykład interakcji między SHRDLU a człowiekiem (jest to skrócona wersja autentycznego dialogu opisanego w książce Winograda 1972: 8–15). Odpowiedzi SHRDLU zaznaczone są kursywą:

- Podnieś czerwony klocek.
- *OK*. [*SHRDLU podnosi klocek*]
- Znajdź klocek wyższy niż ten, który trzymasz, i włóż go do pudełka.
- *Zakładam, że przez 'go' rozumiesz klocek, który trzymam.* [*SHRDLU wkłada właściwy klocek do pudełka*]
- Co jest w pudełku?
- *Niebieska piramida i niebieski klocek.*
- Czy stół może podnosić klocki?
- *Nie*.
- Czy można postawić piramidę na piramidzie?
- *Nie wiem.*
- Postaw piramidę na piramidzie.
- *Nie mogę.* [*SHRDLU próbuje, ale mu się nie udaje*]
- Niebieska piramida jest moja.
- *Rozumiem*.

– Czy jest w pudełku coś, co jest moje?
– *Tak, niebieska piramida*.

Na pierwszy rzut oka dialog ten sprawia wrażenie, jakby rzeczywiście rozmawiało się z maszyną *myślącą*, która *rozumie*, co się do niej mówi. Ale zdanie „*Rozumiem*" w powyższym dialogu nie ma znaczenia przez nas mu przypisywanego. Po otrzymaniu informacji „Niebieska piramida jest moja" SHRDLU po prostu dodaje do swej „wewnętrznej reprezentacji" niebieskiej piramidy informację o jej przynależności do rozmówcy. Kiedy jest o nią pytany, sprawdza, czy ta „wewnętrzna reprezentacja" ma oczekiwaną własność, i stosownie do tego odpowiada. SHRDLU nie rozumie, co znaczy „posiadać"; tak naprawdę SHRDLU *nic* nie rozumie. Najlepiej widać to na przykładzie innego dialogu z SHRDLU, którego autorem jest John Haugeland (cytowany w Copeland 1993: 92):

– Zbuduj wieżę kościelną.
– *Nie wiem, co to jest „wieża kościelna"*.
– Wieża kościelna to dwa klocki sześcienne, jeden na drugim, z piramidą na szczycie.
– *Rozumiem.*
– Zamienię twoją wieżę kościelną na trzy czerwone klocki.
– *Nie rozumiem, co znaczy „zamienię"*.
– „Zamiana" to swobodna wymiana własności.
– *Nie rozumiem, co znaczy „swobodna"*.
– Przykro mi, ale wydawałeś mi się mądrzejszy.
– *Nie rozumiem, co znaczy „mądrzejszy"*.

W drugiej połowie lat 60. zbudowano też pierwszego prawdziwie mobilnego robota o imieniu Shakey (Raphael 1976). W projekcie tym powiązano rozwijane do tej pory niezależnie systemy wizji komputerowej, analizy języka naturalnego i planowania. Po otrzymaniu instrukcji, na przykład: „Przesuń przedmiot *A* z pozycji *B* na pozycję *C* w pokoju *D*", Shakey oceniał swoje położenie, planował trasę dotarcia do pokoju *D*, a po dotarciu tam wykonywał powierzone mu zadanie. Repertuar możliwych działań, które mógł wykonać Shakey, był jednak niewielki: przemieszczanie się z pokoju do pokoju, otwieranie drzwi, zapalanie światła czy przesuwanie przedmiotów. Co więcej, świat, w którym funkcjonował Shakey, był ściśle zdefiniowany: kilka pokojów ze stołami i krzesłami.

Ani SHRDLU, and Shakey nie stały się zalążkami większych projektów. Terry Winograd stwierdził wprost, że SHRDLU to była ślepa uliczka. ARPA (Advanced Research Projects Agency)[4] wycofała się z finansowania prac nad

[4] ARPA finansowała wówczas *gros* badań nad sztuczną inteligencją w USA.

Shakey'em i nikt już nie wrócił do prób skonstruowania robota ogólnego przeznaczenia. Na czym polegał problem? Otóż wszystkie programy stworzone do tej pory w sztucznej inteligencji stosowały się do tzw. mikroświatów (ang. *microworlds*), a więc ściśle zdefiniowanych i dokładnie opisanych wycinków rzeczywistości. Tak więc Logic Theorist dowodził tylko twierdzeń logiki, SHRDLU manipulował tylko klockami na stole, a nawet Shakey – mimo że nazywano go robotem ogólnego przeznaczenia – potrafił poruszać się i wykonywać proste czynności tylko w obrębie kilku pokoi. Wszelkie próby rozciągnięcia zastosowań tych programów na szersze dziedziny, zdefiniowane nie tak ściśle jak logika czy gra w warcaby, kończyły się fiaskiem. Bariery były dwojakiego rodzaju: pierwsza dotyczyła złożoności obliczeniowej, druga – wiedzy potocznej. Warto od razu podkreślić, że żadnej z tych barier nie udaje się pokonać za pomocą doskonalszych algorytmów, bardziej skomplikowanych programów czy szybszych komputerów. Zmiany wymaga – jak się wydaje – sam paradygmat sztucznej inteligencji. Przyjrzyjmy się zatem bliżej każdej z wymienionych barier.

3. Problemy

Zadania, jakie wykonać ma program komputerowy, są zaimplementowane w postaci algorytmów. Efektywność programu zależy zarówno od fizycznych własności komputera (wielkości pamięci, szybkości CPU itd.), jak i ilości operacji, jakie wykonać musi algorytm. Ten drugi czynnik zależy z kolei od ilości danych na wejściu (łatwiej znaleźć maksimum z 10 liczb niż z 1000 liczb) oraz od skomplikowania samego algorytmu (łatwiej znaleźć maksimum z 10 liczb, niż je posortować). Złożoność obliczeniowa to właśnie miara tego skomplikowania, a definiuje się ją po prostu jako funkcję matematyczną $f(n)$, gdzie n jest ilością danych na wejściu. I tak, algorytm dla znajdowania maksimum z n liczb ma złożoność liniową, $f(n) = n$, bo wystarczy raz przejrzeć wszystkie liczby, czyli wykonać n operacji, by znaleźć ich maksimum. Algorytmy sortowania mają złożoność (zależnie od typu algorytmu) $f(n) = n \log(n)$ lub $f(n) = n^2$, bo w tym przypadku należy wykonać $n \log(n)$ lub n^2 operacji porównywania. Algorytmy o podobnej złożoności organizuje się w klasy; mówi się więc o złożoności wielomianowej, gdzie $f(n) = n^x$, bądź wykładniczej $f(n) = x^n$ (w obu przypadkach x jest dowolną stałą). Okazuje się, że właśnie te dwie klasy złożoności algorytmów różnią się *dramatycznie*, jeśli chodzi o ilość operacji, a więc i czas potrzebny do wykonania algorytmu. Weźmy prosty przykład. Niech $x = 3$, czyli funkcja wielomianowa ma postać $f(n) = n^3$, a funkcja wykładnicza $f(n) = 3^n$. Załóżmy, że jedna operacja wykonywana jest w ciągu mikrosekundy i porównajmy wartości tych funkcji dla dwóch różnych war-

tości n (wielkości danych). Czas wykonywania algorytmu o złożoności wielomianowej dla $n = 10$ wynosi 0,001 sekundy, a dla algorytmu o złożoności wykładniczej jest to 0,059 sekundy. Jeśli jednak $n = 60$, to algorytm pierwszy zakończy pracę po 0,216 sekundach, a drugiemu zabierze to $1{,}3 \times 10^{13}$ stuleci.

Algorytmy o złożoności wykładniczej są więc w praktyce nieobliczalne; mogą one być wykorzystywane wyłącznie do rozwiązywania problemów o małej skali (tzn. dla niewielkiego n). Takie właśnie problemy rozwiązywano w mikroświatach: SHRDLU manipulował tylko kilkoma klockami, Shakey, planując swoje działania, miał do dyspozycji tylko kilka czynności. Wykorzystanie tych samych programów, czy też ściślej mówiąc tych samych algorytmów, do rozwiązywania realistycznych problemów jest po prostu niemożliwe. Wydawać by się mogło, że tego dylematu można jednak uniknąć: skoro barierą są algorytmy o złożoności wykładniczej, dlaczego nie zastosować innych, prostszych algorytmów dla rozwiązywania tych samych problemów? Otóż jednym z największych osiągnięć teorii informatyki ostatniego wieku była obserwacja, że prostszych algorytmów dla większości problemów sztucznej inteligencji najprawdopodobniej nie ma! Udało się do dziś zidentyfikować kilkaset problemów, które określa się mianem NP (ang. *non-polynomial*) (Garey i Johnson 1979), a które mają dwie wspólne im własności. Po pierwsze, dla żadnego z tych problemów nie udało się do tej pory znaleźć rozwiązania o prostszej niż wykładnicza złożoności. Po drugie, znalezienie prostego, a więc wielomianowego rozwiązania dla jednego z nich rozwiązuje w ten sam sposób je wszystkie[5]. Dla sztucznej inteligencji był to wynik szczególnie dotkliwy, bo większość problemów, które usiłowano w sztucznej inteligencji rozwiązać, należy właśnie do klasy NP (problem planowania, który rozwiązywał Shakey, okazał się należeć do klasy o złożoności wyższej nawet niż NP). Dla wielu krytyków był to argument za tym, że sztuczna inteligencja oparta na rozwiązaniach algorytmicznych, czyli obliczeniowych, jest niemożliwa.

Drugi poważny problem, który napotkano w sztucznej inteligencji, dotyczył opisu wiedzy potocznej. Jeśli maszyna myśląca ma wchodzić w interakcje z ludźmi i funkcjonować w ich środowisku, musi – choćby częściowo – podzielać ich obraz świata. Wiedza potoczna, która ten obraz świata opisuje, musi być zatem w jakiś sposób reprezentowana w języku maszyny. Prawdopodobnie nikt nie zdawał sobie sprawy, jak trudny to może być problem, aż do czasu publikacji artykułu, w którym McCarthy i Hayes (1969) zdefiniowali tzw. problem ramy (ang. *frame problem*). Autorów interesowało sformalizowanie myślenia zdroworozsądkowego na użytek planowania działań przez robota.

[5] Do tej pory nie znaleziono co prawda *dowodu*, że problemy z klasy NP nie mają rozwiązań wielomianowych, ale znakomita większość badaczy nie wątpi, że taki dowód zostanie przeprowadzony.

Skonstruowany w tym celu tzw. rachunek sytuacji pozwalał opisywać rezultaty działań i przeprowadzać stosowne wnioskowania. Na przykład poniższe zdanie:

(1) (Zachodzi[Na(x,y), S] & Zachodzi[Pusty(x), S]) →
Zachodzi([Pusty(y), Rezultat(<Zdejmij(x,y)>, S)])

charakteryzuje następujące działanie: jeśli w sytuacji S obiekt x jest na obiekcie y i x jest pusty (tzn. nic na nim nie ma), to w rezultacie zdjęcia obiektu x z obiektu y w sytuacji S, obiekt y również będzie pusty. A co z kolorem obiektu y? Czy zmieni się on w wyniku zdjęcia go z x? Dla nas odpowiedź jest jasna: nie. Ale odpowiedź ta nie jest wcale oczywista z punktu widzenia opisywanej teorii. Sytuacja, która powstała w wyniku zdjęcia obiektu x z obiektu y, jest *inną* sytuacją niż sytuacja S. Dopóty, dopóki nie stwierdzimy *explicite*, że kolor obiektów nie zmienia się w wyniku ich przenoszenia, nie wiemy, czy obiekt zachował swój kolor. Problem wydaje się trywialny, ale bynajmniej nie ma trywialnego rozwiązania na gruncie formalizmu logicznego. Dodanie tzw. aksjomatów ramy, a więc twierdzeń typu „obiekt nie zmienia swego koloru w wyniku jego przenoszenia" jest nie do przyjęcia z trzech powodów. Po pierwsze, w warunkach realnego świata takich aksjomatów byłoby nieprzewidywalnie wiele, a mianowicie tyle, ile jest różnych par działanie-własność („obiekt nie zmienia kształtu w wyniku przenoszenia", „obiekt nie zmienia koloru w wyniku przekręcania" itp.). Dodawanie nowych własności i nowych działań do opisu świata wymagałoby nieustannego dodawania takich aksjomatów. Po drugie, prawdziwość takich aksjomatów zależy od kontekstu. Jeśli jeden robot przenosi klocki, a drugi je maluje, to powyższy aksjomat jest fałszywy. Wreszcie, jeśli aksjomaty ramy miałyby opisywać sposób myślenia człowieka w sytuacjach takich, jak opisana powyżej, to na pewno opisują ją fałszywie.

Klasyczny problem ramy można więc sformułować tak oto: jak opisać w zwięzły sposób, co się zmienia, a co pozostaje takie samo w wyniku naszych działań. Problem ramy okazał się jednym z najtrudniejszych do rozwiązania problemów sztucznej inteligencji (Gryz 2013). Napisano na jego temat dziesiątki, jeśli nie setki artykułów naukowych i jemu wyłącznie poświęcono kilka naukowych konferencji (Pylyshyn 1987; Ford, Pylyshyn 1996). Problem ten sprowokował gorące debaty na temat metodologii sztucznej inteligencji i skłonił wielu badaczy do odrzucenia logiki jako języka zdolnego sformalizować myślenie zdroworozsądkowe; żadne z proponowanych rozwiązań problemu nie było bowiem w pełni satysfakcjonujące (Morgenstern 1996). W istocie większość badaczy sądzi, że problem ramy to tylko symptom innego, ogólniejszego problemu i że to *ten* problem powinien być rozwiązany najpierw. „Nie ma

sensu «rozwiązywać» problemu ramy, jeśli oznaczałoby to «od-rozwiązanie» jakiegoś innego problemu" (Janlert 1996). Zanim przejdziemy do próby analizy tego innego problemu, omówmy jeszcze krótko dwa inne problemy zidentyfikowane w kontekście planowania, które są w praktyce nie do rozwiązania na gruncie reprezentacji logicznej.

Pierwszy z nich to problem rozgałęzionych efektów (ang. *ramification problem*) (Finger 1988). Problem dotyczy trudności, a właściwie niemożliwości wyliczenia wszystkich skutków naszych działań. Wyobraźmy sobie, że wchodzimy do biura i stawiamy teczkę koło kaloryfera. W teczce było niestarannie zapakowane śniadanie, które pod wpływem ciepła zaczyna rozpływać się pośród znajdujących się w teczce notatek. Były to notatki przygotowane na referat habilitacyjny, który mamy wygłosić za pół godziny. Notatki nie dają się odczytać, nasz referat kończy się klapą, rada wydziału odmawia nam habilitacji. Takie opóźnione czy odległe konsekwencje naszych działań są bardzo trudne do zidentyfikowania i opisania.

Okazuje się, że nie tylko skutków, ale również warunków naszych działań nie jesteśmy w stanie wyliczyć. Klasyczny przykład to próba uruchomienia samochodu. Spodziewamy się, że wystarczy do tego przekręcić kluczyk w stacyjce. Okazuje się, że to nieprawda: akumulator musi być naładowany, w baku musi być benzyna, rura wydechowa nie może być zatkana kartoflem, nikt w nocy nie ukradł nam silnika itp. Trudno oczekiwać, że bylibyśmy w stanie podać *wszystkie* warunki niezbędne do podjęcia jakiegokolwiek działania. Które z nich powinniśmy zatem wyspecyfikować dla zdefiniowania działań dostępnych dla robota? To problem uszczegółowienia warunków (ang. *qualification problem*) (McCarthy 1986).

Wszystkie trzy wymienione wyżej problemy pojawiły się w kontekście prób kodyfikacji wiedzy potocznej. Do tej pory próba ta w sztucznej inteligencji się nie powiodła[6]. Jack Copeland wymienia trzy warunki, jakie powinna spełniać taka kodyfikacja (Copeland 1993: 91). Po pierwsze, po to, żeby wiedza potoczna była dla maszyny myślącej użyteczna, musi być przechowywana w zorganizowany sposób. Trudno sobie wyobrazić, że cała dostępna baza danych będzie sprawdzana za każdym razem, kiedy maszynie potrzebna jest jedna konkretna informacja. Po drugie, aktualizacja tej wiedzy musi odbywać się niezwykle sprawnie, zarówno pod względem czasu, jak i poprawności. Przesunięcie krzesła nie powinno wywoływać lawiny wnioskowań na temat możliwych zmian jego własności (poza położeniem), ale powinno uruchomić natychmiastową aktualizację stanu szklanki z herbatą na nim stojącej. Wresz-

[6] W 1984 roku zainicjowano ogromny projekt CYC, którego celem było stworzenie bazy danych zawierającej dużą część naszej wiedzy potocznej (Lenat i Feigenbaum 1991). Nie rozwiązywało to jednak w żaden sposób problemów dyskutowanych powyżej.

cie, maszyna postawiona przed konkretnym problemem do rozwiązania musi być w stanie określić, jaka część dostępnej jej wiedzy jest istotna dla rozwiązania problemu. Brak reakcji silnika na przekręcenie kluczyka w stacyjce wymaga być może sprawdzenia stanu akumulatora, ale nie obecnej fazy księżyca.

Jak zatem my, ludzie, przechowujemy i wykorzystujemy wiedzę potoczną? Otóż nie wiemy tego, bo w ogromnej większości przypadków świadome rozumowanie jest tu nieobecne. Stosowna wiedza dociera do naszej świadomości *tylko* wtedy, gdy popełniliśmy błąd (przesunęliśmy krzesło ze szklanką wylewając przy tym herbatę) albo gdy próbujemy zaplanować coś nietypowego (wnieść pianino po wąskich schodach). Ale nawet wtedy introspekcja jest bezużyteczna, bo nic nam nie mówi na temat mechanizmów, jakich wówczas używamy. Wedle Daniela Dennetta (1987) wielkim odkryciem sztucznej inteligencji było spostrzeżenie, ze robot (komputer) to słynna *tabula rasa*. Żeby maszyna myśląca mogła funkcjonować w rzeczywistym świecie, musi mieć całą potrzebną jej wiedzę podaną *explicite*. Co więcej, wiedza ta musi być tak zorganizowana, by zapewnić maszynie sprawny do niej dostęp. Wyzwanie, jakie stoi przed sztuczną inteligencją, można więc sformułować tak oto: w jaki sposób, używając formalizmu logicznego, uchwycić holistyczną, otwartą i wrażliwą na kontekst naturę wiedzy potocznej? Do tej pory zupełnie nie wiemy, jak to zrobić.

4. Rozpad

Najsurowsza krytyka sztucznej inteligencji nadeszła ze strony filozofii, głównie za sprawą Johna Searle'a i Huberta Lederera Dreyfusa. Słynny argument „Chińskiego Pokoju" Searle'a (1980) zdawał się pokazywać, że niemożliwe jest, aby program komputerowy cokolwiek „rozumiał" czy też miał „świadomość". Tym samym program budowy maszyny *myślącej* przy użyciu programów komputerowych jest skazany na niepowodzenie. Argument Searle'a jest w istocie argumentem przeciwko funkcjonalizmowi i komputacjonizmowi w filozofii umysłu, a uderza w sztuczną inteligencję, bo ta przyjęła obie te tezy filozoficzne w swoich założeniach. Tymczasem krytyka Dreyfusa skierowana jest wprost przeciwko sztucznej inteligencji i podważa jej pozostałe, być może nawet istotniejsze założenia. I tak Dreyfus zarzuca badaczom sztucznej inteligencji, że bez większej refleksji „przejęli od Hobbesa tezę, że rozumowanie to liczenie, od Kartezjusza reprezentacje umysłowe, od Leibniza ideę «prawd pierwotnych» – zbioru podstawowych elementów, za pomocą których można wyrazić całą wiedzę, od Kanta tezę, że pojęcia to zasady, od Fregego formalizację takich zasad i od Wittgensteina postulat atomów logicznych z *Trakta-*

tu" (Dreyfus 1992: 1137). Ten bagaż filozoficzny można wyłożyć w postaci następujących założeń, które według Dreyfusa poczyniła sztuczna inteligencja (Dreyfus 1992):

- Założenie biologiczne: mózg to mechanizm manipulujący symbolami, podobnie jak komputer.
- Założenie psychologiczne: umysł to mechanizm manipulujący symbolami, podobnie jak komputer.
- Założenie epistemologiczne: inteligentne zachowanie można sformalizować i tym samym skopiować przy użyciu maszyny.
- Założenie ontologiczne: świat składa się z niezależnych, nieciągłych faktów.

Dreyfus odrzuca wszystkie te założenia[7], korzystając z argumentów wziętych z filozofii Martina Heideggera i Maurice'a Merleau-Ponty'ego. Wskazuje trzy aspekty inteligentnych zachowań, które jego zdaniem zostały całkowicie pominięte w sztucznej inteligencji. Po pierwsze, większość naszych codziennych zachowań wykonywana jest w oparciu o umiejętności, czyli *wiedzę, jak*, a nie uświadomioną *wiedzę, że*. Po drugie, ciało jest integralnym elementem inteligentnych zachowań; nasza zdolność stawiania czoła nowym sytuacjom wymaga ulokowanego w tych sytuacjach fizycznego ciała. Po trzecie, zachowanie ludzkie nie jest nastawione po prostu na cel, ale na wartości, i jako takie zawsze zależy od kontekstu czy sytuacji. Dreyfus przekonująco argumentuje, że żaden z tych aspektów ludzkiej inteligencji nie może być skopiowany, czy nawet symulowany, w tradycyjnej sztucznej inteligencji.

Dreyfus miał status swego rodzaju *enfant terrible* w sztucznej inteligencji i w pewnym momencie był niemal powszechnie ignorowany. Nie pomogło mu na pewno odwoływanie się do fenomenologii (dla badaczy o nastawieniu inżynierskim to był po prostu bełkot) ani skrajność w wyrażaniu poglądów: „[Sztuczna inteligencja] to paradygmatyczny przypadek tego, co filozofowie nauki nazywają wyrodniejącym programem badawczym" (Dreyfus 1992: IX). Z drugiej strony, wiele diagnoz Dreyfusa się sprawdziło. W szczególności jego podkreślenie roli ciała i kontekstu (sytuacji) dla wyjaśniania inteligentnych zachowań zostało w późniejszym okresie z sukcesem wykorzystane (Brooks 1991).

Lata 80. przyniosły pierwsze próby komercjalizacji rozwiązań dokonywanych w sztucznej inteligencji. Jedną z najbardziej popularnych aplikacji były tzw. systemy eksperckie, rozwijane w oparciu o udane prototypy akademickie: DENDRAL (Feigenbaum, Buchanan, Lederberg 1971) i MYCIN (Shortliffe

[7] Wydaje się, że trudno byłoby dziś znaleźć badaczy, którzy zgadzaliby się z założeniem biologicznym. Z drugiej strony, bez założenia ontologicznego opis świata przy pomocy języka byłby chyba niemożliwy.

1976). Celem pierwszego z nich było odgadywanie struktury molekularnej próbek chemicznych w oparciu o dane ze spektrometru. Największym wyzwaniem przy budowie tego systemu było sformalizowanie wiedzy posiadanej przez chemików-praktyków. Okazało się bowiem, że podręcznikowa wiedza z chemii tu nie wystarczała: „Szukaliśmy nie tylko twardych faktów z dziedziny chemii o wartościowości czy stabilności (...) procesów, ale także miękkich faktów: jak konkretny naukowiec podejmuje konkretną decyzję, kiedy nie jest pewien, kiedy dostępne są różne świadectwa i gdy sporo jest niejasności" (Feigenbaum w: McCorduck 2004: 329). MYACIN, zaprojektowany w celu diagnozowania infekcji krwi, musiał poradzić sobie w dwójnasób z problemami napotykanymi w poprzednim projekcie. Po pierwsze, nie istniał model teoretyczny opisujący przyczyny i symptomy infekcji krwi, który pozwalałby formułować reguły wnioskowania w tej dziedzinie. Po drugie, reguły takie musiały uwzględniać niepewność i względny charakter świadectw dostępnych w każdym konkretnym przypadku. Mimo tych przeszkód, MYACIN okazał się dużym sukcesem, bo jego diagnozy były znacznie lepsze niż dokonywane przez niedoświadczonych lekarzy i często równie dobre jak dokonywane przez ekspertów.

Komercyjne systemy eksperckie budowane w latach 80. już takimi sukcesami nie były. Zdobycie, sformalizowanie, a następnie właściwe wykorzystanie specjalistycznej wiedzy w jakiejś dziedzinie było problemem nie tylko technicznym, ale dotykało znowu problemu wiedzy potocznej. Po raz kolejny obietnice dokonywane w sztucznej inteligencji, tym razem w wydaniu komercyjnym, nie sprawdziły się. Okres od końca lat 80., charakteryzujący się utratą wiary, ale także zainteresowania sztuczną inteligencją, często nazywany jest „zimą" tej dziedziny. Wtedy też sztuczną inteligencję ogarnął trwający aż do dziś kryzys tożsamości. Dla większości badaczy stało się oczywiste, że postęp, jaki dokonał się w ciągu 40 lat przy projekcie zbudowania maszyny myślącej, a więc takiej, która przeszłaby test Turinga, był znikomy. Ogrom pracy, funduszy i ludzkich talentów zaangażowanych w tę dziedzinę był niewspółmierny do osiągniętych efektów. Co więcej, wielu badaczy miało poczucie, że droga, którą obrano, prowadzi donikąd. Obrazowo podsumował to Stuart Dreyfus: „Dzisiejsze zapewnienia i nadzieje na postęp (...) w dziedzinie budowy inteligentnych maszyn są jak przekonanie, że ktoś, kto wszedł na drzewo, dokonał postępu w drodze na Księżyc" (Dreyfus, Dreyfus 1986: 10). Z drugiej strony, nawet najzagorzalsi krytycy programu sztucznej inteligencji musieli przyznać, że w niemal każdej z jej *pod*dziedzin dokonania były znaczące. Nic zatem dziwnego, że to właśnie w tych poddziedzinach koncentruje się dziś wysiłek badaczy. Począwszy od lat 90., ze sztucznej inteligencji wyodrębniła się robotyka, wizja komputerowa, przetwarzanie języka naturalnego, teoria decyzji i wiele innych. Każda z tych dziedzin organizuje swoje własne

konferencje i wydaje pisma naukowe, w niewielkim tylko stopniu wchodząc w interakcje z innymi dziedzinami. Tak oto, dla większości badaczy, podniosły cel zbudowania maszyny myślącej stał się celem ulotnym czy wręcz nieistotnym.

Czy jest zatem coś, co spaja dziedzinę nazywaną wiąż jeszcze sztuczną inteligencją? Jaki cel przyświeca tym wszystkim, którzy usiłują zbudować systemy percepcji, podejmowania decyzji, wnioskowania czy uczenia się? Wydaje się, że wyróżnić można co najmniej trzy stanowiska w tej sprawie. Pierwszą postawę reprezentują ci, którzy nadal uważają, że celem sztucznej inteligencji jest zbudowanie maszyny imitującej zachowanie (a więc i myślenie) człowieka. Nazywają ten nurt sztuczną inteligencją na poziomie ludzkim (ang. *human-level AI*). Choć należą do tego nurtu nestorzy sztucznej inteligencji: McCarthy (2007), Nilsson (2005) czy Minsky, to niewątpliwie stanowią oni dziś mniejszość wśród badaczy. Druga postawa apeluje o odrzucenie ludzkiego zachowania i myślenia jako wzorca dla sztucznej inteligencji: „Naukowym celem sztucznej inteligencji jest zrozumienie inteligencji jako metody obliczeniowej, a jej celem inżynierskim jest zbudowanie maszyn, które przewyższają ludzkie zdolności umysłowe w jakiś użyteczny sposób" (Ford, Hayes 1998). Dla uzasadnienia skuteczności tej metody badacze przywołują przykład aeronautyki, która odniosła sukces dopiero wówczas, gdy przestała w konstrukcji samolotów imitować ptaki. Trzecia wreszcie postawa, najbardziej, jak się wydaje, powszechna, to całkowite porzucenie idei budowy maszyny *myślącej*. Zamiast tego proponuje się budowę „racjonalnych agentów" (ang. *rational agents*), a więc czegoś, co zachowuje się w taki sposób, który maksymalizuje zadaną miarę wydajności. „Standard racjonalności jest matematycznie dobrze zdefiniowany. (...) z drugiej strony, ludzkie zachowanie jest dobrze dostosowane do jednego tylko typu środowiska, a zdefiniowane jest, no cóż, przez wszystko, co ludzie robią" (Russell, Norvig 2010: 5). Obrońcy tego podejścia konsekwentnie sugerują więc zmianę nazwy dziedziny ze „sztucznej inteligencji" na „racjonalność obliczeniową".

Druga z opisanych wyżej postaw jest najbliższa programowi kognitywistyki. Badamy procesy poznawcze – ludzkie, zwierzęce czy maszynowe – nie po to, by zbudować imitację człowieka, ale po to, by go zrozumieć. Wykorzystujemy w tym celu dane z każdej nauki, która nam na ten temat coś może powiedzieć: psychologii, neurofizjologii, lingwistyki czy właśnie sztucznej inteligencji w ostatnim rozumieniu. Projektowanie robotów nie musi być już celem samym w sobie, ale środkiem do zrozumienia zachowań i procesów myślowych ludzi i zwierząt (Webb 2008). Wielu badaczy obarcza test Turinga odpowiedzialnością za wyznaczenie nieosiągalnego, niepotrzebnego i niewiele znaczącego wzorca dla sztucznej inteligencji.

> Test [Turinga] doprowadził do powszechnego niezrozumienia właściwych celów naszej dziedziny. (...) Zamiast ograniczać zakres badań (...) do prób imitacji ludzkich zachowań, możemy osiągnąć dużo więcej badając, jak zorganizowane są systemy obliczeniowe, które wykazują się inteligencją (Ford, Hayes 1998).

Wydaje się więc, że HAL[8] na długo pozostanie tylko filmową fantazją.

Bibliografia

Brooks R. (1991), *Intelligence Without Representation*, „Artificial Intelligence" 47, s. 139–159.

Campbell M., Hoane A.J. Jr., Hsu F.-H. (2002), *Deep Blue*, „Artificial Intelligence" 134, no. 1–2 (January), s. 57–83.

Copeland J. (1993), *Artificial Intelligence: A Philosophical Introduction*, Blackwell.

Darrach B. (1970), *Meet Shakey, The First Electronic Person*, „Life", 20 November 1970.

Dennett D. (1987), *Cognitive Wheels: The Frame Problem of AI*, w: Pylyshyn (ed.) 1987, s. 41–64.

Dietrich E. (ed.) (2012), *Philosophy of Artificial Intelligence*, http://philpapers.org/browse/philosophy-of-artificial-intelligence.

Dreyfus H.L. (1992), *What Computers Still Can't Do*, MIT Press.

Dreyfus H.L. (2007), *Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian*, „Artificial Intelligence" 171, nr 18, s. 1137–1160.

Dreyfus H.L., Dreyfus S.E. (1986), *Mind Over Machine*, New York: Macmillan/Free Press.

Feigenbaum E.A., Buchanan B.G., Lederberg J. (1971), *On Generality and Problem Solving: A Case Study Using the DENDRAL Program*, „Machine Intelligence", Vol. 6, Edinburgh University Press.

Ferrucci D. i in. (2010), *Building Watson: An Overview of the DeepQA Project*, „AI Magazine" 31, nr 3, s. 59–79.

Finger J.J. (1988), *Exploiting Constraints in Design Synthesis*, Technical Report STAN-CS-88-1204, Computer Science, Stanford University.

Ford K.M., Hayes P.J. (1998), *On Computational Wings: Rethinking the Goals of Artificial Intelligence*, „Scientific American Presents" 9, no. 4 (Winter), s. 78–83.

[8] HAL musiał uznać wyższość człowieka. Czytelnik, który nie potrafi zidentyfikować HAL-a, musi uznać wyższość maszyny. Program Watson (Ferrucci i in. 2010), który zwyciężył w popularnym teleturnieju „Jeopardy", bez trudu odgadłby jego tożsamość.

Ford K.M., Pylyshyn Z.W. (1996), *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*, Westport, CT: Ablex Publishing Corporation.

Garey M.R., Johnson D.S. (1979), *Computers and Intractability*, W.H. Freeman.

Gelernter H. (1959), *Realization of a Geometry-Theorem Proving Machine*, International Conference on Information Processing, Paris, s. 273–282.

Goldstine H. (1972), *The Computers from Pascal to von Neumann*, Princeton, NJ: Princeton University Press.

Gryz J. (2013), *The Frame Problem in Artificial Intelligence and Philosophy*, „Filozofia Nauki", R. XXI, nr 2 (82), s. 15–30.

Janlert L.-E. (1996), *The Frame Problem: Freedom or Stability? With Pictures We Can Have Both*, w: Ford, Pylyshyn 1996, s. 35–48.

Lenat D.B., Feigenbaum E.A. (1991), *On the Tresholds of Knowledge*, „Artificial Intelligence" 47, s. 185–230.

McCarthy J. (1986), *Applications of Circumscription to Formalizing Commonsense Knowledge*, „Artificial Intelligence" 28, s. 86–116.

McCarthy J. (2007), *From Here to Human-Level AI*, „Artificial Intelligence" 171, s. 1174–1182.

McCarthy J., Hayes P. (1969), *Some Philosophical Problems from the Standpoint of Artificial Intelligence*, w: B. Meltzer, D. Mitchie (eds), *Machine Intelligence*, Edinburgh: Edinburgh University Press, s. 463–502.

McCorduck P. (2004), *Machines Who Think*, 2nd ed., A.K. Peters Ltd.

Morgenstern L. (1996), *The Problem with Solutions to the Frame Problem*, w: Ford, Pylyshyn 1996, s. 99–133.

Newell A., Simon H.A. (1956), *The Logic Theory Machine*, „IRE Transactions on Information Theory" (September).

Nilsson N.J. (2005), *Human-Level Artificial Intelligence? Be Serious!*, „AI Magazine" 26, nr 4, s. 68–75.

Pylyshyn Z.W. (1984), *Computation and Cognition: Towards a Foundation for Cognitive Science*, Cambridge: MIT Press.

Pylyshyn Z.W. (ed.) (1987), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex Publishing Corporation.

Raphael B. (1976), *The Thinking Computer: Mind Inside Matter*, San Francisco: W.H. Freeman.

Russell S.J., Norvig P. (2010), *Artificial Intelligence. A Modern Approach*, Prentice Hall.

Samuel A.L. (1959), *Some Studies in Machine Learning Using the Game of Checkers*, „IBM Journal of Research and Development" 3, no. 3, s. 210–229.

Searle J. (1980), *Minds, Brains and Programs*, „Behavioral and Brain Sciences" 3, nr 3, s. 417–457.

Shortliffe E.H. (1976), *Computer-Based Medical Consultations: MYCIN*, Elsevier/North-Holland.
Simon H.A. (1965), *The Shape of Automation: For Man and Management*, New York: Harper & Row.
Simon H.A., Newell A. (1958), *Heuristic Problem Solving: The Next Advance in Operations Research*, „Operations Research" 6, s. 1–10.
Simon H.A., Newell A. (1976), *Computer Science as Empirical Inquiry: Symbols and Search*, „Communications of the ACM" 19, no. 3, s. 113–126.
Thagard P. (2008), *Cognitive Science*, w: *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2008/entries/cognitive-science/.
Turing A. (1969), *Intelligent Machinery*, „Machine Intelligence", Vol. 5, Edinburgh: Edinburgh University Press.
Webb B. (2008), *Using Robots to Understand Animal Behavior*, „Advances in the Study of Behavior" 38, s. 1–58.
Winograd T.A. (1972), *Understanding Natural Language*, Academic Press.

Streszczenie

Sztuczna inteligencja pojawiła się jako dziedzina badawcza ponad 60 lat temu. Po spektakularnych sukcesach na początku jej istnienia oczekiwano pojawienia się maszyn myślących w ciągu kilku lat. Prognoza ta zupełnie się nie sprawdziła. Nie dość, że maszyny myślącej dotąd nie zbudowano, to nie ma zgodności wśród naukowców, czym taka maszyna miałaby się charakteryzować, ani nawet, czy warto ją w ogóle budować. W artykule starałem się prześledzić dyskusję metodologiczną towarzyszącą sztucznej inteligencji od początku jej istnienia i określić relację między sztuczną inteligencją a kognitywistyką.